THE ELEMENTS OF STATISTICAL METHOD



THE MACMILLAN COMPANY
NEW YORK · BOSTON · CHICAGO · DALLAS
ATLANTA · SAN FRANCISCO

MACMILLAN AND CO., LIMITED
LONDON · BOMBAY · CALCUTTA · MADRAS
MELBOURNE

THE MACMILLAN COMPANY
OF CANADA, LIMITED
TORONTO

THE ELEMENTS OF STATISTICAL METHOD

BY

WILLFORD I. KING, M.A.

INSTRUCTOR IN STATISTICS IN THE UNIVERSITY OF WISCONSIN

New York

THE MACMILLAN COMPANY

LONDON · MACMILLAN & CO., LTD.

Set up and electrotyped Published April, 1912.

· PRINTED IN THE UNITED STATES OF AMERICA ·

DEDICATED

TO

THOMAS SEWALL ADAMS, PH.D.

OF THE

WISCONSIN TAX COMMISSION

PREFACE.

The purpose of this book is to furnish a simple text in statistical method for the benefit of those students, economists, administrative officials, writers, or other members of the educated public who desire a general knowledge of the more elementary processes involved in the scientific study, analysis, and use of large masses of numerical data. While it is intended primarily for the use of those interested in sociology, political economy, or administration, the general principles set forth are applicable likewise to every variety of statistical data. The author has found that the members of his classes in this subject were not, as a rule, expert mathematicians, and he believes that this is true of a majority of those persons who are called upon to make practical use of statistics, hence, no pretense whatever has been made, in this work, of presenting any but the most simple of the mathematical theorems upon which statistical method is based.

So far as the author is aware, there is no book published in America which attempts to cover the field of statistical method in its present state of advancement. There are several excellent treatises published abroad but they either embrace but a part of the subject or are adapted especially to the biologist, to the advanced

student of statistics, or to those having considerable mathematical training. Under these circumstances, it is believed that there is place for an elementary text of this nature.

References are given to only a few of the principal works on the subject. It is not expected that the student will read all of those listed at the close of any chapter, but, when all are available, it will usually be best to make use of them in the order in which they are named. If more advanced study of any topic is desired, the student will find abundant references cited in those books.

My thanks are due to Dr. Horace Secrist, Professor John R. Commons, and Professor T. K. Urdahl, all of the University of Wisconsin, for reading my manuscript and offering me many valuable suggestions which have resulted in its improvement.

To Dr. Thos. S. Adams, my former instructor, now of the Wisconsin Tax Commission, I am indebted for the major part of all that has made this work possible, and any merit which it may possess must be credited largely to his efforts.

WILLFORD I. KING.

UNIVERSITY OF WISCONSIN,
September, 1911.

TABLE OF CONTENTS.

PART I. INTRODUCTION.

PART II. THE GATHERING OF MATERIAL.

PART IV. COMPARISON OF VARIABLES.

APPENDICES.

LIST OF DIAGRAMS.

THE ELEMENTS OF STATISTICAL METHOD

PART I.

INTRODUCTION.

CHAPTER I.

THE HISTORICAL DEVELOPMENT OF STATISTICAL SCIENCE.

Sec. 1. **Preliminary Remarks.**

During the last fifty years, the proper system of dealing with large numbers, in other words the **science** of statistics, has first attained importance in the public mind and has first been perfected to such an extent as to really have attained the dignity of a science. Nevertheless, this stage is but the latest story to be added to a building whose foundations were laid many centuries since, and, while the main purpose of this book is to deal with the methods of procedure used in the existing stage of advancement, it seems that the present picture may stand out more clearly if placed in its proper setting by means of briefly outlining the development of statistics from its early beginnings down to the present day.

Sec. 2. **Statistics in Ancient Times.**

The growth of statistics is coeval with the growth of national organization. As soon as tribes were united into coherent confederacies or distinct nations, it be-

came necessary for the ruler to collect facts concerning his domain. He must needs know something of its wealth in order to compute the amount of taxes or tribute which he might levy. He must have a census of his fighting men so that he might know his war strength. On certain occasions, unusual situations might arise which would necessitate enumerations for other specific purposes. One of our earliest statistical compilations tells us of the collection of data concerning the population and wealth of Egypt in order to make arrangements for the construction of the pyramids. This occurred about 3050 B.C.[1] Many centuries later (about 1400 B.C.), Rameses II took a census[2] of all the lands of Egypt in order to reapportion them among his subjects in such a way as he deemed proper.

In the first and second chapters of the Book of Numbers, we read how Moses numbered the tribes of Israel doubtless with a view of determining their fighting strength. Another census was taken by David, about 1018 B.C., for similar purposes.[3] Even in the Far East, similar forces were at work and it is recorded that the Chinese government had a description of the provinces compiled by Yuking as early as 1200 B.C.

Herodotus tells us that "Lykurgus divided the territory of Laconia into 39,000 portions, assigning to the

[1] Herodotus II, 109.

[2] Herodotus II, 177.

[3] II Samuel XXIV.

Spartans 9,000 portions and to the Lacedaemonians 30,000 portions."[1]

This was but one of many enumerations made by the ancient Greeks for the purposes of apportioning land, levying taxes, classifying the inhabitants and determining the military strength. In Rome, after the days of Servius Tullius, quite elaborate censuses were taken, primarily for the purposes of taxation and ascertainment of population, and the inhabitants of the city were required to register births and deaths at certain specified temples.

During the Middle Ages, the feudal barons and their imperial over-lords frequently enumerated the population and property of their domains, investigations of this nature by Charlemagne, William the Conqueror, Al-Mamum, Emperor Frederick II of Germany, and Edward II of England being among those recorded by historians of the day.

It should be noted that, in each of the above instances, the intent of the census was to aid the government in its administrative work or its plans for war; taxation, land distribution, and available soldiers being the commonest subjects of inquiry calling forth an enumeration. With the exception of the Roman censuses, such inquiries seem to have been undertaken only when some special reason existed for collecting the data and not at any regular intervals.

[1] Meitzen, A., *Statistics*, p. 16.

Sec. 3. **Mercantilistic Period.**

During the period in which Mercantilism dominated the policies of the Western European governments, a marked increase in the bulk of statistics collected is noticeable. This was the natural outcome of the belief that the government should encourage certain lines of industry and should adopt all necessary measures to secure a favorable balance of trade. In order to judge correctly of needs for and effects of the various kinds of legislation, more elaborate statistics were necessary than had hitherto been considered essential. Besides, the growth of centralized monarchy, with the accompanying elaboration of government, gave a greater necessity for extensive statistical information than had been required during the Middle Ages, and with this greater necessity, came also the increase of ability required to successfully carry out such investigations as might be desired. Success now attended that monarch who could, in advance, best measure his resources as compared with those of his rivals, and then best husband these resources for times of conflict.

We find Philip II of Spain making extensive inquiries in 1575 A.D., from the prelates and corregidors of Spain concerning the districts over which they had supervision. At the beginning of the seventeenth century, Sully prepared for his master, Henry of Navarre, a comprehensive statement of the financial and military resources of France; in 1665, Colbert com-

piled extensive statistics of trade;[1] and, in 1699, Louis XIV required reports on the state of the country from each of the general intendants.

It was Prussia, however, which, in modern times,[2] first began a systematic, periodic collection of statistical data. In 1719, Frederick William I began gathering semi-annual reports as to population, occupations, houses, real estate holdings, taxes, city finances, etc. At a later date, these figures were collected and tabulated at intervals of only three years. Frederick the Great, likewise, was a firm believer in the value of statistical information and he enlarged the scope of the inquiries by including such things as nationality, age, deaths and their causes, data concerning agriculture, trade, manufactures, shipping, value of property, etc. He even took a decided personal interest in the work and, as a result, the accuracy and completeness of the information was immensely improved. During the period 1747 to 1782, a complete statistical system was thus worked out.

Sec. 4. **The Modern Census.**

The idea of the decennial census seems to be an American product. The provision for representation in the lower house of Congress in accordance with population made a census indispensable, and hence, this

[1] Block, M., *Traité de Statistique*, p. 25, also Bertillon, *Cours Élémentaire de Statistique*, p. 27.

[2] Meitzen, A., *Statistics*, p. 27.

was provided for in the Constitution and the first census was taken in 1790. Slightly over a decade later, (in 1801), England, too, adopted a similar plan of enumeration.

The German Zollverein of 1833, which eliminated interstate duties within the German boundaries and preserved only a common external customs barrier, provided for a distribution of the proceeds of the tariff according to population.[1] In order to secure a correct apportionment, a triennial census was established. The idea of a regular enumeration gained steadily in favor and was adopted by one after another of the civilized nations. Finally, in 1911, we find China taking her first official census.

As time has passed, the censuses have grown larger and larger in scope, and during the last three or four decades, have become extremely elaborate. In 1900, the United States established a permanent census bureau which devotes itself continuously to the working out of many special problems and the elucidation of the statistics collected during the regular census periods.

Most leading nations also have special statistical bureaus which, by means of scientific estimates, attempt to keep the statistics of a nation abreast of the times. An example of this idea in the United States is our national Bureau of Statistics. Many of the states have also provided similar bureaus for their respective needs.

[1] Bertillon, J., *Cours Élémentaire de Statistique*, p. 23.

Sec. 5. Comparative Statistics.

Today we have great masses of statistics collected by numerous sources, public and private, but with little co-ordination of work as regards different bureaus in the same nation. As a result, while each collection is valuable for purposes of its own, comparison of data for different places is still very difficult. In the earlier statistical inquiries, this was probably not even thought of but, with the appearance of national rivalry among the leading European nations and the general adoption of a Mercantilistic policy, such comparisons began to be made.

As early as 1544, Sebastian Muenster, a professor at Heidelberg, published a systematic treatise on the ancient countries, their organization, wealth, armies and fighting strength, commerce, church relations, laws, etc.[1] In 1562, Francesco Sansovino,[2] and, in 1589, Giovanni Botero, both Italians, published works of a similar nature, and, in 1614, Pierre d'Avity, Seigneur de Montmarin, followed with a more accurate and complete treatise in four volumes dealing, likewise, largely with comparative statistics. These works were patterns for numerous others which followed at later dates until, today, we have statistical dictionaries treating of almost every conceivable field of investigation, but still necessarily deficient in accuracy because of the

[1] Meitzen, A., *Statistics*, p. 20.

[2] John, V., *Geschichte der Statistik*, p. 38.

lack of uniform inquiries in the various parts of the world.

Sec. 6. **Vital and Social Statistics.**

We have outlined briefly the growth of statistics gathered by the various governments in order to measure their strength or assist in their administration but, in the early part of the seventeenth century, certain new uses were suggested for some of the statistical data which had been collected. During the early period of the Reformation, the Protestant churches, largely in an effort to check illegitimacy, required the registration in the church records of all births, deaths, and marriages. In many German and English cities these decrees were carried out with a fair degree of completeness. In 1612, Professor George Obrecht,[1] of Strasburg University, proposed that the government keep a complete record of all vital and criminal statistics, worked out a plan for carrying out his ideas, and illustrated, with remarkable insight, the uses to which such statistics could be put in devising methods for reforming the morals of the people, and also for providing a system of life insurance and pensions.

In 1661, Capt. John Graunt, of London, made the first recorded analytical study in the field of vital statistics.[2] He came to the conclusions that the birth and death rates were quite constant; that the births

[1] Meitzen, A., *Statistics*, p. 25.

[2] John, V., *Geschichte der Statistik*, p. 225 f.

were distributed among the sexes in the ratio of 14 boys to 13 girls; that the deaths among a given hundred of persons born could be calculated for each succeeding year; and, therefore, that, from an accurate birth record, the total population of the country could be computed.

In 1691, Caspar Neumann,[1] prebendary of Breslau, collected, from the parish registers of that city, records of 5,869 deaths and, from these figures, proceeded to demonstrate that no such fateful significance as had usually been supposed could be attached to the ages seven or nine. His notes and conclusions fell into possession of the Royal Society of England, and thus came to the attention of the noted astronomer and scientist, Edmund Halley. He utilized Neumann's figures in the computation of the first recorded complete life table and derived therefrom the expectation of life at each age and a scientific system of life insurance though he failed to take into account, in his calculations, the increase in population.

Life insurance was, at this time, by no means unknown, having originated in wagers on the death of the captain of a vessel, these wagers being used to protect the ship owners in case of the loss of the vessel. A similar system is still in vogue as regards the life of some prominent person whose death might seriously inconvenience business. This system, however,

[1] John, V., *Geschichte der Statistik*, p. 208 f.

was entirely unsuited to the person who wished to insure his own life for the benefit of his family and the charges had never been, in any degree, scientifically exact. Hence, Halley's tables formed the foundation for life insurance in the modern sense. In 1698, the first life insurance institution was founded in London and, a year later, the "Society of Assurancy for Widows and Orphans" came into existence.[1]

Statistics was further allied to the province of mathematics by Jacques Bernouilli, a professor of Basel, who died in 1705, leaving behind a work in which he mathematically eludicated the theory of probabilities, a theory which has played no small part in the development of modern statistical science. Another advance was made by Johann Peter Süssmilch[2] who, in 1741, published a treatise in which he attempted to demonstrate, statistically, the doctrine of the "Natural Order."[3] He showed the approximate equality in numbers of the sexes at the time of marriage and construed this as a divine command in favor of monogamy. He, further, worked out the age constituency of the population and the constancy of the ratio between births and deaths. The fact that the death rate was larger in the city than in the country he interpreted to mean that in the cities luxury and vice flourished, hence bringing down the wrath of God.

[1] Meitzen, A., *Statistics*, p. 33.

[2] Meitzen, A., *Statistics*, p. 35.

[3] John, V.. *Geschichte der Statistik*, p. 269 f.

Süssmilch's statistical studies were followed up during the early part of the nineteenth century by those of the renowned scientists Laplace and Fourier. A little later, Lambert Adolphe Jacques Quetelet, a prominent Belgian astronomer and mathematician, made extensive statistical studies in the realms of astronomy and meteorology. His investigations concerning the weather led him to a study of the periodical phenomena of vegetation and it was but a step further to include the animal kingdom and then mankind in the scope of his research. The social and moral as well as the physical characteristics of men were embraced in the field of inquiry. He made the surprising discovery that similar results were obtained from each and every variety of phenomena observed. In each case, a certain mean or norm was found to exist[1] about which the number of instances or occurrences was great, and as the distance from the mean increased the number of items fell off with mathematical regularity. In fact, he found that if the numbers of occurrences were plotted as ordinates that the result was a regular binomial curve identical with that given by the mathematical law of chance or probability. This seemed to indicate that man's actions are governed wholly by this same law for he showed that all kinds of human acts occurred with

[1] Quetelet described in detail the characteristics of the mean or average man representing the normal type of the race.

marked regularity. Crimes, suicides, accidents, all showed comparatively constant figures. This, he believed, proved man to be largely the product of his environment, society to be responsible for the individual, yet he expressly denied any restraining force on the individual and rejected fatalism as an explanation.[1]

Many of Quetelet's followers, however, continued his reasoning to what they believed to be a logical conclusion. This is exemplified by Sir F. W. Herschel who, in 1850, declared that man was wholly the creature of environment and that free will, if existent, was practically non-perceptible. H. Thomas Buckle, the historian, voiced his approval of the same idea.

The Italian school, on the other hand, were not at all ready to accept this doctrine without question and they were supported by many German statisticians. In 1871, Gustav Schmoller[2] illuminated the question by asserting that the regularity of human action was only due to regularly acting causes which, in general, tended to produce constant results. Free will was shown by the fact that results were not entirely regular but, at times, varied decidedly from the usual order, even though the causes remained constant. This view has since gained wide acceptance.

[1] Meitzen, A., *Statistics*, p. 75, also John, V., *Geschichte der Statistik*, p. 332 f.

[2] Meitzen, A., *Statistics*, p. 88.

Sec. 7. **Statistics as an Aid to Economics.**

From the earliest times, statistics were considered necessary as an aid to administration and, in the days of the Mercantilists and Cameralists, many governmental policies were based on statistical information. Late in the seventeenth century, Gregory King had attempted to show statistically a fixed relationship between the supply and price of commodities. Most of the economic writers of the eighteenth century made more or less use of numerical data to establish their theories, but it remained for the Historical School of economists to emphasize the importance of statistics in the economic field. Since they assumed that economic laws and doctrines were not to be reasoned out abstractly but proven historically and concretely, statistics became, to them, a prime necessity. Bruno Hildebrand, who had much practical experience as a governmental statistician, was a leading expositor of this idea and the correct methods of applying statistics were greatly developed by his contemporary, Karl Knies.[1]

Sec. 8. **Statistical Method.**

In analyzing statistical data, it was, of course, discovered, even in the earliest experiments, that some definite method must be followed in order to render the results intelligible. As data became more abundant and many new fields were opened up to

[1] Meitzen, A., *Statistics*, p. 77

investigation, this subject of method became more and more complex. Refined scientific inquiries could not be conducted by the crude and cumbersome machinery suited only to simpler problems. As early as 1741, Anchersen,[1] a Dane, devised statistical tables for the comparisons of European states[2] and, in 1782, Crome, of Giessen, went further and utilized geometric figures for like purposes. F. J. Mone, in his "Theory of Statistics," 1824, emphasized the necessity of carefully worked out methods for the solution of all statistical problems. Ernst Engel, in his "Methods of Enumerating Population," 1861,[3] developed this idea more completely than ever before and he was ably seconded by Rumelin, Knies, Adolf Wagner, and M. Block. Within the last three or four decades, the development of the pure theory of statistics has had a remarkable growth. Such men as August Meitzen, Francis Edgeworth, Francis Galton, E. L. Thorndike, Karl Pearson, G. Udny Yule and C. B. Davenport in the field of biological statistics, and Jacques Bertillon, Arthur L. Bowley, R H. Hooker, Thos. S. Adams, and Warren Persons in the field of economics have aided in advancing the theory far beyond its former bounds. It is to the simpler outlines of this branch of statistics that this volume is devoted.

[1] John, V., *Geschichte der Statistik*, p. 88.

[2] Meitzen, A., *Statistics*, p. 41.

[3] Meitzen, A., *Statistics*, p. 96.

Sec. 9. Instruction in Statistics.

The early lecturers and writers who included among their teachings the subject which we now call statistics only introduced numerical quantities as a detail in the instruction in general political, geographic, and economic information and it was to this entire field of thought that the term "statistics" was first applied. The first recorded lectures in this line were given by Hermann Conring at the University of Helmstedt in 1660.[1] Conring was a noted physician and also a professor of natural law. His statistical lectures dealt, however, with data concerning the land and its products and the state and its resources. The Cameralists of the early eighteenth century followed his example and made "statistics" a part of their teachings. The first, however, to organize this mass of knowledge into a logical whole was Gottfried Achenwall (often called the "Father of Statistics") a professor of the University of Marburg. It was he who first applied to this line of study the appellation "statistics," deriving the term from the Italian word "statista" meaning statesman. Achenwall's main idea was the comparison of one state with another in order to find a correct guide for political action. He dealt in his lectures with Spain, Portugal, France, Great Britain, the Netherlands, Russia, Denmark, and Sweden.[2] He began his lectures in 1746.

[1] Meitzen, A., *Statistics*, p. 22, and John, V., *Geschichte der Statistik*, p. 52.

[2] Meitzen, A., *Statistics*, p. 24, also John, V., *Geschichte der Statistik*, p 74 f.

We have noted the fact that, at this time, the instruction given under the title of "statistics" included all that the schools then taught of political economy and geography. Adam Smith first really segregated political economy as a separate science when he published "The Wealth of Nations" and this work was soon followed by the teachings and writings of Stewart, Malthus, Ricardo, Say, Sartorious, Jacob, and Kraus.

In the early part of the nineteenth century, C. Ritter published his "Science of the Earth in Relation to the Nature and History of Men." This voluminous work tended to establish geography, also, as an independent science. Life insurance, which, as we have seen, played an important role in the origin of statistical studies likewise tended, as it was perfected, to become a distinct branch of study.

Only in comparatively recent years have courses been offered in the leading universities of Europe and America on the science of statistics as the term is now limited and defined, and, even yet, the study is almost invariably taken up as a subordinate branch of political economy or biology.

Sec. 10. **Different Branches of Statistics.**

The modern domain of statistics can be generally divided into two main provinces, **statistical method** and **applied statistics.**

Statistical method may properly be considered a branch of mathematics, inasmuch as it attempts to

formulate definite rules of procedure applicable in handling groups of data of many different varieties. Many of these rules apply equally well to economic or biological data while, in other cases, special rules may be developed which are best adapted to one particular field. The methodologist is not concerned in any but the most indirect way with the specific investigations for which the general laws, rules, and methods which he formulates are to be used.

Applied statistics, as the name signifies, consists of the application of the rules and formulae laid down by the methodologist to the concrete facts as they exist, the relationship being the same as in the case of pure and applied science. While the methodologist is likely to be primarily a mathematician, the specialist in applied statistics may be a census expert, a state official, a sociologist or philanthropist, a biologist, an economist, an insurance actuary, or an investigator in almost any other branch of human knowledge.

The province of applied statistics may be legitimately subdivided again into two general fields — **descriptive** and **scientific.** The **descriptive** field deals with records, either of things in their existing state or from the historical standpoint. The U. S. census with its manifold tables comparing the people and resources of the different sections of the United States both for the given time and with preceding decades is a most excellent example of well-developed descriptive statis-

tics. In this branch of statistics, nearly every citizen is interested, since everyone wishes to know whether his state or nation is growing in population or wealth, whether foreigners are more or less numerous than in the neighboring commonwealths, etc.

For the scientist, however, statistics have still a different interest. He wishes to use the data which record past events in order to establish definite physical or psychological laws. The biologist is anxious to verify his hypothesis concerning heredity, the meteorologist wishes to trace a connection between sun-spots and temperature, the economist desires to verify the quantity theory of money, and the statesman endeavors to demonstrate the salutary effects of a tariff law Thus, **scientific** statistics makes use both of the rules laid down in statistical method and the data collected for descriptive purposes. Scientific statistics, then, is the ultimate goal toward which a large part of the work of all modern statisticians ultimately tends.

Sec. 11. **Summary.**

We have thus traced very briefly the development of the science of statistics from its primitive form to its present complex status. With this preliminary setting, the student will be perhaps prepared better to take up the study of statistical method as set forth in later pages

REFERENCES.

MEITZEN, AUGUST. *History, Theory and Technique of Statistics.* Translated by Roland P. Falkner. Amer. Acad. of Pol. and Soc. Science, Phila., 1891. Part I.

JOHN, V. *Geschichte der Statistik.* Ferdinand Enke, Stuttgart, 1884.

BLOCK, MAURICE. *Traité Theorique et Pratique de Statistique.* Guillaumin et Cie, Paris, 1886. Chaps. I and II.

BERTILLON, JACQUES. *Cours Élémentaire de Statistique.* Société d'Editions Scientifiques, Paris, 1895. Première Partie.

YULE, G. UDNY. *An Introduction to the Theory of Statistics.* J. B. Lippincott and Co., Phila., 1911. Introduction.

CHAPTER II.

THE SCIENCE DEFINED.

Sec. 12. Definition of Statistics.

We have seen, in the preceding chapter, that many different forms of knowledge have, in the past, been termed statistics and that, today, the science has shifted far from its old meaning, the study of the state. Our next task is to formulate a definition suited to the science in its present-day aspect, a definition which is inclusive enough to take in all that is still classed under this nomenclature and exclusive enough to keep out all ideas extraneous to that title. Webster defines the term "statistics" thus: "Classified facts respecting the condition of the people in a state . . . especially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement." This definition accords with the etymology of the word, which it will be remembered is derived from *statisi* and more remotely from *state*. Until comparatively recent times, this phase of statistics was almost the only one worthy of mention, but nowadays, the term is applied to the study of data obtained in the fields of biology, astronomy, etc., so that the definition must, necessarily, be expanded to cover the new uses.

A similar idea, but somewhat more comprehensive

is expressed by the following definition given by Bowley:[1] "Statistics is the science of the measurement of the social organism, regarded as a whole, in all its manifestations." This statement, however, as its author says, limits the science to only one field — that of man and his activities. Modern statistics takes into consideration, however, biological, astronomical, and physical as well as social phenomena, hence, the definition is obviously too narrow.

Statistics has also been denominated "the science of counting."[2] This obviates the above-mentioned error of confining the definition to only one field. On the other hand, serious defects of another kind at once appear in this definition. The major part of statistical work includes not only mere counting but also the further process of making estimates. In collecting its statistics of wheat grown in the United States, the Department of Agriculture does not attempt to get an actual record of each bushel produced but simply obtains estimates of the present year's crop as compared with that of previous seasons. In this way, crop reports of a fair degree of accuracy can be obtained with almost no actual counting. In fact, in dealing with large numbers, an accurate count is almost always a physical impossibility. It is self-evident, for example, that a large number of people are not counted when a

[1] Bowley, A. L., *Elements of Statistics*, p. 7.

[2] Bowley, A. L., *Elements of Statistics*, p. 3.

census is taken and the names of many others are entered in the lists of two or more different enumerators.

Another defect of this definition is that it would seem to apply only to the collection of data and not to the analysis of the material collected while, as a matter of fact, both parts are essential to any complex statistical study. Hence, this definition must also be rejected as inadequate.

One of the prime objects of statistics is to give us a bird's-eye view of a large mass of facts, to simplify this extensive and complex array of isolated instances and reduce it to a form which will be comprehensible to the ordinary mind. To attain this end, averages are very often used, hence, Bowley says: "Statistics may rightly be called the science of averages."[1] But modern statistics goes further than to present mere averages. By graphic processes, we see **variation,** that is, the fluctuations with regard to a certain standard or norm, portrayed as, for example, when we chart the temperature changes of a season or a cycle. By pictograms are shown relative totals, familiar illustrations being the bars or squares used to show the relative population, wealth, products, or expenditures of different nations. By correlation tables and coefficients, relationships are indicated. Therefore, this definition, likewise, seems far too restricted.

A more comprehensive definition, and the one which

[1] Bowley, A. L., *Elements of Statistics*, p. 7.

we shall adopt for the purposes of this work, is the following: **The science of statistics is the method of judging collective natural or social phenomena from the results obtained by the analysis of an enumeration or collection of estimates.** This is certainly more inclusive than any of the preceding definitions and, while it is possible that statistical problems might be imagined which would not fall within its limits, it is sufficiently broad for practical purposes.

REFERENCES.

BOWLEY, ARTHUR L. *Elements of Statistics.* Chas. Scribner's Sons, N. Y., 1907. Chap. I.

MEITZEN, AUGUST. *Statistics*, pp. 89–109

YULE, G. U. *Introduction to Statistics*, Introduction.

BLOCK, MAURICE. *Traité de Statistique*, Chap. IV.

CHAPTER III.

USES, CHARACTERISTICS AND SOURCES OF STATISTICS.

Sec. 13. Necessity of Statistical Science.

The human mind is so constituted that it cannot image and comprehend a large number of distinct impressions at any one time. As a result, it is impossible to compare intelligently two complex groups of things without simplification of the groups in some manner. It would be a man of exceptional mnemonic power who could, after listening to the reading of two lists of one hundred items each stating the names and wealth of the respective inhabitants of two villages, give any intelligent opinion as to the comparative riches of the two communities. If this is true for such small groups as this, it evidently would be utterly impossible to make comparisons of the wealth of great nations without some manner of reducing the mass of separate facts to a simple whole. The same would, of course, be true in the case of any other phenomena involving large numbers. What could one understand of the amount of lumber contained in a forest from a description of the separate trees? How could one compare the climates of different localities by a study of their daily weather records? It is for the purpose of simplifying these unwieldy masses of facts that statistical science

is useful. It reduces them to numerical totals or averages which may be abstractly handled like any other mere numbers. It draws pictures and diagrams to illustrate general tendencies and, thus, in many ways adapts these groups of ideas to the capacity of our intellects.

Sec. 14. **Uses of Statistics.**

The facts having been once simplified they are now in a shape where they may be used for purposes of comparison and this is one of the principal aims which the science of statistics has in view. We are interested to know the population of the United States not, primarily, for the value of that fact in itself, but, principally, in order that we may compare the population of today with that of past decades and thus picture in our minds the nation's growth or that we may compare the numbers of our people with the numbers of other lands or that we may compare the growth in our population with the growth of our food supply, our manufacturing or mining industries, our increase in wealth, or any one of many similar things. Thus, it is **relative** rather than **absolute** size which appeals to our imaginations.

These comparisons, however, are seldom made simply with the idea of satisfying our idle curiosity. They are necessary in order to settle the most weighty questions of government and economics. How could congressional constituencies be justly apportioned without a

population census? Is tuberculosis increasing or decreasing? The answer to this question must be shown statistically and has weighty import from the points of view of public finance and general public policy in fighting the disease just as well as from the standpoint of the health of the inhabitants. Should the railroads be allowed to increase their freight rates? Before this can be determined, we must have reliable statistics of earnings and expenditures. In fact, few important actions can be properly taken by a modern government, or even a modern corporation, without a statistical study of conditions in the field in question.

The rapid growth of the science of accounting and the general demand for uniform systems of accounts for all municipalities has greatly emphasized the need of correct statistical methods in this line.

Numerous commissions have recently sprung into being whose duties vary from the government of cities to the regulation or investigation of almost every phase of private or governmental activity. The recommendations or decrees of these commissions must, in nearly all cases, rest largely upon statistical information and the merit of the results obtained therefore depends primarily on the correctness of the statistical method employed and the accuracy with which it is carried out.

Every insurance company must base its rates upon computations derived through the study of large masses of data. Since new forms of insurance are constantly

being evolved and since conditions of life are ever changing, new statistics must continually be collected and new calculations as continuously be made. An example of this is the insurance of workingmen against unemployment, a question which is now first being seriously considered in many countries. As yet, the statistics are far too incomplete to make possible a comprehensive scientific system of insurance in this line even if proper methods and safeguards can be worked out.

But practical statesmen and men of affairs are not the only ones who find in statistics a most valuable ally. The theoretical economist or the scientific investigator in the field of natural phenomena must likewise constantly call on statistics for proof or verification of his hypotheses or theories. The biologist thereby verifies the laws of variation and heredity. The economist seeks to establish laws of population, of wages, of prices, or to show the connection between different groups of phenomena as, for example, financial crises and unemployment. The sociologist would demonstrate the relationship of sales of alcoholic liquor to crime, poverty, suicide, and similar phenomena where some connection exists or is suspected.

Bowley says: "The proper function of statistics, indeed, is to enlarge individual experience."[1] Without a statistical study, most of our ideas are likely to be

[1] Bowley, A. L., *Elements of Statistics*, p. 8.

decidedly vague and indefinite. A reduction to figures gives clear cut form to this hazy conception, enables us to set objects in their proper perspectives and relationship, and, hence, gradually to work out the laws governing their movements and changes.

Sec. 15. **Law of Statistical Regularity.**

One of the most valuable characteristics of modern scientific statistics is that it succeeds in giving us a sufficiently accurate picture of a group of objects without going through the laborious and expensive process of a complete enumeration of all the items in the group. Thus, it is by no means necessary in ascertaining the average wage of American workingmen to obtain data regarding each man at work. If certain typical instances can be obtained and properly averaged, the difference from the true average wage of all the working men is likely to be such a small quantity as to be, for all practical purposes, negligible. Similarly, the anthropologist can discover the physical characteristics of a tribe or race by taking careful measurements of only a small minority of the whole. This is due to the law of nature formulated in the mathematical theory of probabilities which states that a moderately large number of items chosen **at random** from among a very large group are almost sure, on the average, to have the characteristics of the larger group. Thus, if two persons, blindfolded, were to pick here and there three hundred walnuts each from a bin containing a million

nuts, the average weight of the nuts picked out by each person would be almost identical even though the nuts varied considerably in size. Furthermore, if one were to obtain the average weight of the whole million it would not differ, essentially, from the average weight of either of the smaller groups.

This principle may be easily verified by taking a small number of dice and throwing them forty or fifty times. If four dice are taken, the total number of spots on both sides is twenty-eight. On the average, half of these, or fourteen, should turn up each time. In fifty throws, the total number of spots turned up should be 700. Experiment will show that the approach to this number will be surprisingly close. It is upon this principle that gamblers are enabled to run continuously and profitably with only small odds in their favor. It is this same principle which gives rise to the regularity in the number of crimes and the number of suicides, facts which, as we have seen, once greatly troubled the advocates of the doctrine of Free Will. It is this principle which makes possible insurance against death or other calamities. This principle is frequently denominated **the law of statistical regularity.**

It must not, however, be inferred from the above that any number of samples, no matter how large, will give **exactly** the same results as would be obtained by the use of the entire mass of data. The probability of

error diminishes constantly as the number of items used as samples increases. If, then, only a few sample items are used the chance error is likely to be so large as to seriously vitiate the results but, as the number of samples chosen grows large, the error diminishes until it eventually becomes negligible.

Sec. 16. **Inertia of Large Numbers.**

The **law of inertia of large numbers** is a corollary of the law of statistical regularity. It arises from the fact that, in most classes of phenomena, when one part of a large group is varying in one direction, the probabilities are that another equal part of the same group is varying in the opposite direction; hence, the total change will be slight. Thus, for example, while the amount of wheat produced in any one locality varies immensely from year to year, the wheat production of the world, as a whole, remains relatively stable for decades. The losses from fire in a single city may be fifty times as large in a given year as in the preceding one but the annual losses throughout the entire country will remain almost constant. Hence, a fire insurance company can, years in advance, calculate its losses with a fair degree of accuracy. Statistical science, then, is largely based on the theory of probabilities and its corollaries.

This property of inertia by no means precludes the possibility of change with the passage of time. It only means that when the numbers involved are of great

magnitude the change is likely to be more regular than in those cases in which small quantities are involved. Thus, while the fire losses in the United States would be relatively constant from year to year they might nevertheless be steadily diminishing owing to a tendency to erect stone or concrete buildings instead of frame structures. Similarly, the wheat production of the world gradually increases as new lands are brought into cultivation.

This property of inertia is less evident when, for any reason, there is greater probability of variation in one direction than in the other. If, for example, in a given state practically all the cities had borrowed up to the legal debt-limit, a curtailment of the debt of one city would rarely be offset by an increase in the debt of some other, hence the stability of the total of the city indebtedness within the state would be affected to a larger extent relatively by such decreases in the debts of individual cities than if the maximum debt-limit were non-existent.

Sec. 17. **Distrust of Statistics.**

It is said that non-scientific people may be divided into two classes as regards their attitude toward new inventions or discoveries. One class accepts without question the wildest stories of incredibly marvelous discoveries and wonders why no one stumbled upon them before. The other class, usually possessing a little more education, are skeptical of all scientific truth

and label it all alike as "guess-work." Similar attitudes of mind, in regard to statistics, are noticeable among those unfamiliar with that science. The attitude of the first class is well expressed by the old proverb "Figures won't lie," while the other class, a little more erudite, are prone to characterize all statistics as tissues of falsehood. Either theory can be readily proven by selecting proper examples.

One of the shortcomings of statistics is that they do not always bear on their face the label of their quality. The crudest table, founded on the most unreliable basis, appears, to the casual observer, equally valuable with a table compiled after months of labor by a corps of skilful statisticians. To judge of the value of a statistical presentation, it is, then, usually essential to know something of the author and his reliability and skill as a statistician. Yet, to a careful observer, the internal evidence in the table itself may be a valuable clue as to its merit. An amusing example of failure to observe intelligently is found in the book of a recent writer on socialism, the main thesis of which is based on an erroneous table taken from a government report, the errors in the table being so glaring as to be at once evident to anyone in the least familiar with statistical data.

It is true that one can prove anything by statistics but he can only do so by unscientific handling of his data or deliberate manipulation of the figures with the purpose of showing the desired result. It is also a fact

that some sets of figures are hard to analyze and their meaning is often doubtful and debatable but, in a large percentage of cases, a clear and indisputable result may be arrived at if the analysis is conducted in a scientific and unbiased manner. The science of statistics, then, is a most useful servant, but only of great value to those who understand its proper use."

Sec. 18. **Progressive Accuracy in Statistics.**

As we have seen, the value of statistics depends primarily on the accuracy of the figures. Accurate figures are often very hard indeed to obtain. Is there any excuse, then, for the preparation of any statistical table or statement based on figures of doubtful accuracy or reliability? As a matter of fact, such tables and statements have marked scientific worth though scientific fairness should always lead the author to accompany the table with a statement concerning his sources of information, and the likelihood of accuracy or error therein, in order that his reader may not be misled. The value of such investigations is chiefly in the fact that they may serve as foundations for future work of a more accurate nature. Every preliminary investigation shows up the difficulties to be overcome, the weakness or strength of methods used and some general ideas as to the probable result. Most detailed investigations are difficult, if not impossible, without some such preliminary work. The measurements of the velocity of light only attained their present accuracy through a long series of approxi-

mations. The perfection of modern life tables was only attained by a succession of improvements on the crude tables drawn up by Casper Neumann from the church records. Therefore, inaccurate investigation is only to be condemned when its results are stated, not as a preliminary or tentative basis for further work, but as a final and definite conclusion.

Sec. 19. **Limitations of Statistics.**

Statistics, while an extremely useful tool to the investigator in almost any line of scientific inquiry, has its limitations and shortcomings which cannot be overcome. Statistics largely deals with averages and these averages may be made up of individual items radically different from each other. In the average, these irregularities are all swallowed up. Methods of analysis have been devised, as we shall see later, which partially obviate this defect but no system which makes a large and complex group intelligible to the mind at a glance can avoid effacing most of the minor irregularities. We usually assume, and in general correctly, that these items are of no importance, but this assumption is not always in accordance with the facts. It may be true that the match industry includes but an insignificant fraction of our working population and it may also be true that but a small fraction of these workers suffer from phosphorus poisoning but, for those affiicted, the fact that their numbers do not affect appreciably the general average does not lessen their torture, does not

appear to them to be and is not a sufficient reason for foregoing legislation to remedy the evil. Statistics, from the very nature of the subject, cannot and never will be able to take into account individual cases. When these are important, other means must be used for their study.

Sec. 20. **Sources of Statistical Information.**

There are several ways in which the statistician may proceed to collect the necessary data for his work. The first method is that of individually collecting data for himself. In this case, of course, it is usually only possible to take samples of the mass of items to be analyzed but the investigator has the advantage of knowing exactly the conditions of the investigation and the accuracy of sampling is entirely within his own hands. This method is largely used in the natural sciences and, to some extent, in sociological and economic investigations but is usually impossible in the latter lines if the field to be covered is extensive.

A modification of this method is for the investigator to employ a number of enumerators or correspondents and compile the results of their counts or estimates. This has actually been done by the large speculators on the Chicago Board of Trade in order to get advance information as to crop conditions of the world. Evidently, this procedure entails a heavy expense and is beyond the means of most scientific inquirers.

A second mode of obtaining information is to turn to

the enumerations and estimates made by other private statisticians, compare their work, and compile the results. This is difficult and usually impracticable because of lack of information as to the sources of their material, and the methods and accuracy of their inquiries and also because the different investigators did not cover similar fields or have similar ends in view

As a result of these difficulties, most of the numerical research work in the fields of economics or sociology must proceed by aid of government reports. It is only the government which usually has both the means and the inclination to collect unbiased statistics concerning its subjects and their activities throughout the country as a whole. During the last few decades, the variety of data gathered by the Census and other Statistical Bureaus of the different nations has been so extensive and varied as to afford a mine of valuable material for the patient investigator in the field of the social sciences. An additional advantage which the government has over individuals in carrying on its inquiries is that it may use compulsion in obtaining information. Naturally, this can in no way insure that the answers to questions will be truthful but it helps in overcoming the inertia and negligence of the informants, two of the most serious obstacles in private investigations.

Sec. 21. **Phases of Statistics.**

As we have already noticed in studying the historical development of the subject, there are several

different phases of statistics each of which is or has been emphasized by certain schools at certain times. Historically, the development was as follows:

I. Empirical Statistics.

(*A*) Used principally as aid to administration

II. Comparative Statistics.

(*A*) Used as a basis of economic doctrine.

(*B*) State policies based thereon.

III. Analysis of Statistics by Scientific Methods.

(*A*) Statistics now better adapted to verify economic, social, and scientific hypotheses and theories.

(*B*) Value as a guide to governmental action greatly enhanced

At the present time, two distinct branches of statistics may be discerned.

I. Statistical Method.

II. Statistical Information.

A knowledge of the first of these is essential for the statistician in order that he may correctly obtain the second. The latter is the branch which is of interest to the general public. To use a simile from the field of engineering, the public cares nothing for the course of mathematics which the engineer must pursue in order to correctly construct a great bridge. It is interested only in the results. To the engineer, however, the mathematics is of prime importance. To attempt to handle statistics properly without a

knowledge of statistical method is only a little less absurd, though vastly more common, than to attempt to build great steel bridges without a knowledge of trigonometry. This volume is primarily devoted to the study of those elementary methods, a knowledge of which is essential to all desiring to engage in statistical work at first hand, especially in the realm of the social sciences.

REFERENCES.

MEITZEN, AUGUST. *Statistics*, pp. 55–110, 143–155, 207–219.

BOWLEY, A. L. *Elements of Statistics*, Chap. I.

BLOCK, MAURICE. *Traité de Statistique*, Chap. IV.

BERTILLON, JACQUES. *Cours Élémentaire de Statistique*, Parts IV and V.

BOWLEY, ARTHUR L. *An Elementary Manual of Statistics.* Macdonald and Evans, London, 1910. Chap. I.

PART II.

THE GATHERING OF MATERIAL.

CHAPTER IV.

THE PROBLEM TO BE SOLVED.

Sec. 22. **Defining the Problem.**

The first thing upon which the statistical investigator, when beginning his work, must decide is the exact nature of the problem which he desires to solve. Even a slight change in its scope or form may require an entirely or partially different method of procedure. If, for illustration, a person wishes to begin a study of comparative wages in order to demonstrate some general theory or proposition, he must first decide as to whether the requirements of his problem demand a knowledge of money wages or real wages. Next, he must be sure as to whether he needs to know the wages paid for a definite amount of effort, for making a certain product, or for working a certain length of time, or whether the inquiry relates to the income of the working man himself per year or to the total income of the man and his family for the same period. Each of these problems is a distinct one and would require entirely different methods of determination. **The first essential then is to make the problem definite and clear-cut.**

Sec. 23. Selection of Factors of Problem.

In almost every statistical problem, some comparison is involved and this comparison can usually be expressed as a percentage or a ratio. If we wish to compare death rates in various cities, we usually speak of it as a certain number per thousand of inhabitants; in referring to growth of cities, we use percentages while, in giving a comparative record of suicides, we would probably state the results as a certain number per hundred thousand. In either of these cases, a numerator and denominator are required and the quotient may be referred to as a coefficient. It is of the utmost importance that the terms of this fraction be selected with the greatest care for, if these are erroneously chosen, the final comparison will be vitiated if not rendered worthless. It is this variety of error which is most subtle and most likely to deceive the investigator himself and the results are so specious that the public naturally accept them at their face value and later, when their falsity is demonstrated, another impetus is given to the feeling of distrust toward statistics in general.

An example of such a fallacy, due to the use of erroneous factors, was furnished by a newspaper in a discussion of the American navy during the Spanish-American war. It was stated that the death-rate in the navy during the war period was only nine per thousand while in the city of New York for the same

period the death-rate was sixteen per thousand. The conclusion was drawn that it was safer to be a sailor in our navy in war time than to live in New York City. A little reflection, however, will convince one that such a conclusion is not warranted by the figures given. In obtaining this ratio, the total number of deaths was taken as the numerator in each case and the denominators were respectively the total number of persons living in New York City and the total number of sailors in the navy. But, as a matter of fact, these numbers were wholly incomparable. It is a well known fact that the death rate is very high among young children and among old people. But the personnel of the navy is composed almost wholly of young men in the prime of strength and vigor. Not only this, but each must pass a strict examination to show that he is healthy and robust. Thus, the weak and diseased are eliminated. Evidently, the facts would require that the death rate in the navy be compared with the death rate of a similar picked body of men in New York City before any legitimate conclusions could be drawn regarding the comparative chances of death in the two places.

Similarly, if one is desirous of comparing the number of murders in the Klondike with the number in Chicago in order to draw conclusions concerning the general propensity of the people of each place toward violence, it would be entirely improper to divide the number of

murders in each locality by the respective populations. Murders are, in the great majority of cases, committed neither by women, children nor the old and decrepit. Again, while the numerators have been correctly selected, the denominators should be approximately the respective numbers of the male population between the ages of sixteen and sixty in each place.

Bertillon, the eminent French statistician, gives the following rule for coefficients—"Always compare effects to the causes producing them."[1] This idea may be restated thus—**Be careful to so select the quantities used as numerator and denominator, in each case, that the quotients derived may be legitimately compared.**

REFERENCES.

MEITZEN, AUGUST. *Statistics*, pp. 168–170.

[1] Bertillon, Jacques, *Cours Elementaire de Statistique*, p. 94.

CHAPTER V.

THE STATISTICAL UNIT.

Sec. 24. Determining the Unit.

The science of statistics deals with numbers and numbers pre-suppose units. At first glance, the determination of a unit seems a very simple matter indeed but, more often, the opposite is true. Before a definite conclusion can be arrived at, it is usually necessary to take into consideration the nature of the result desired.

As we have seen, one of the oldest forms of inquiry is the counting of the people. In the United States, we have the decennial census whose original purpose was to determine the population of each of the states in order that representatives might be correctly apportioned among the same. To this end, the Constitution provides, in Amendment XIV, Sec. 2, "Representatives shall be apportioned among the several States according to their respective numbers, excluding Indians not taxed." The question arises as to whether or not the makers of the Constitution intended this to be taken literally. For purposes of Congressional apportionment, would a **person** include an Indian laboring in a city if his name did not appear on the tax roll or does it refer only to Indians on reservations? Would the term include the half-breed on the reservation? If so, does it take in the person of one sixteenth Indian blood

or, indeed, the "squaw man"? Will this unit—the person—include the French traveller who happens to be in New York on the day of the census? If it refers only to residents, will it count the Dakota farmer who is doing his week's trading across the border in Canada when the enumerator appears on the scene? If he is included, how about his son who is with the engineer corps in Panama for a few months? These few questions will show the difficulty of defining such a simple unit as **a person** according to the requirements of our Constitution.

The Census Bureau attempts to find the number of farms at each decade. How is the term to be defined? Is or is not a five acre market garden a farm? Does the term include the thousands of acres of government land ranged over by the cattle of a ranchman owning a single quarter section? If a man owns two eighty-acre tracts half-a-mile apart and works them both, do they constitute one or two farms? If his hired man lives on one eighty does that change the status? If he rents one eighty to a tenant what is the effect? If he has a large plantation with a dozen tenants thereon all under his superintendence, how many farms are there? These difficulties are better illustrated in the introduction to Volume I on Agriculture in the U. S. Census of 1900, pp. xiii to xvii.

But if such units as **a person** or **a farm** are hard to define explicitly, what about the difficulty when such a

unit as **a criminal** is in question? To be a **criminal** must one be guilty of felony? What about the man who commits murder but bribes the jury to acquit him? Manifestly, in this case, we arrive at obstacles that are all but insurmountable, yet, a definition of a criminal seems indispensable before one can obtain any comparative statistics of crime.

That the unit as finally decided upon shall correspond to the name given it is highly desirable, but, even if the suitableness of the name is in doubt, **it is not only desirable but strictly essential that the unit be accurately and unmistakably defined and that the same unit be used in each of the periods or places between which it is intended to make comparisons.** In order that the unit shall be perfectly explicit, it is necessary that its definition shall, **before beginning the investigation, be worked out in minute detail** so as to cover all imaginable questions which may arise concerning it. If the person defining the unit does not intend to conduct the investigation himself but expects to employ enumerators, it is essential that the **unit be described in such clear terms** and that **the details of the definition be so conveniently outlined and arranged** that the enumerator can easily find and comprehend every sentence of the instructions. Enumerators are of only ordinary intelligence and, if any considerable number are employed, the least ambiguity is certain to **give** rise to confusion

Sec. 25. Necessary Characteristics of the Unit.

Not only must the unit selected be defined with precision but **it must also be of such a nature that it may be correctly ascertained.** Suppose that one is interested in determining the comparative education in two communities. It would be absurd to select as the unit for the numerator **the educated person,** for it would be impossible to define a unit that would fit the title and then locate the persons corresponding to the definition. Some simpler unit must be used for the numerator based on some tangible measurement such as a college degree, a certificate of graduation from a high school, a certain number of years' schooling, ability to read and write, familiarity with a certain brief list of facts, or some other specific evidence rather than on the inner characteristics which we think of when we refer to a man as educated. In brief, then, **the abstract must be measured by its concrete manifestations.**

REFERENCES.

Bowley, A. L. *Elements of Statistics*, Chap. III.
Meitzen, August. *Statistics*, pp. 117–118.

CHAPTER VI.

PLANNING THE COLLECTION OF DATA.

Sec. 26. **Preliminary Plans.**

Before the actual collection of material is begun, every phase of the question should be carefully studied in order that no energy should be wasted, errors reduced to a minimum, and the necessity for a second inquiry be avoided. In fact, one of the peculiarities of statistical work is that practically everything must be anticipated in advance, all possible sources of error detected and guarded against, and even the general results estimated. Problems, factors, units, questions, schedules, enumerators, tabulation, methods of work, time, expense, etc., are among the items that must be carefully gone over in minute detail. Statistical work is tedious, at best, and errors and misunderstanding are likely to occur in spite of all precautions, but each hour spent in carefully prearranging the work is likely to save a score of hours in trying to straighten out the confusion due to a hasty and ill-advised program.

As has been said before, several methods of investigation are possible which may be broadly classed under two general heads, primary and secondary.

I. Secondary Investigation.

Sec. 27. **Characteristics.**

For this sort of an investigation, the preliminary work that can be done is slight. Nearly everything

depends on the material collected and plans for the collection are therefore difficult to formulate.

II. Primary Investigation.

Sec. 28. General Characteristics.

In the case of a primary investigation, circumstances are radically different for, in this, the director of the work can make such plans for collection as he believes best adapted to the required end. Four general plans are possible, personal investigation, estimates from correspondents, schedules to be filled by the informants, and schedules in charge of enumerators. The proper method is, of course, determined by the nature of the problem, the accuracy of results desired, and the financial resources available.

Sec. 29. Personal Investigation.

This method is especially adapted to intensive studies. A good example of work along this line is that conducted by Le Play, in Europe. He studied workingmen's budgets by spending several months in the home of a single family of working people and repeating the process with a number of families. By following this method for many years, he obtained statistics of great accuracy but the number of families which it is possible to study in this way, in a reasonable amount of time, is too small to constitute a fair sample of the whole. Le Play, even in a lifetime, could not study a very large number of families, but it is note-

worthy that large scale investigations conducted since that time have not overthrown the fundamental principles which he set forth. Arthur Young, by his travels, gives us another variety of personal study of a less intensive type. Booth's great work on the "Life and Labour of the People of London" is also partially a personal work and the same may be said of Rowntree's studies in York, England, in 1900, but these are on a less intensive basis than those of Le Play.

This type of inquiry, while admirable because of additional accuracy due to personal supervision, must needs cover too narrow a field to be representative and is also liable to too large an injection of the personal element. The prejudices and desires of the investigator become too often unconsciously woven into the fabric of his conclusions.

Sec. 30. **Estimates from Correspondents.**

When it is desired to obtain only an approximate result, this method is often used because of its ease and inexpensiveness. As has been said, this is a favorite method of obtaining crop reports, estimates usually being stated as a percentage of increase or decrease from the normal or from the preceding year. While individual reports are necessarily quite inaccurate, the errors involved tend to compensate each other and, when a large number of reports are returned, the net results are likely to be approximately correct. A modi-

fication of this plan is that in which agents are sent through the country to collect the estimates.

Sec. 31. **Schedules to be filled by Informants.**

This is another extensive method and differs from the preceding only in that the questions asked are those concerning which the informant is presumed to have definite, accurate, knowledge. Like any method relying on correspondents, it has the serious defect of depending for its success on persons whose interest in the work is, to say the least, not acute. A large percentage of schedules are usually not returned unless they emanate from the state or some of its representatives endowed with, and actually exercising, compulsory powers. Those schedules that are returned are often extremely incomplete and full of errors. If they are very simple, the probabilities of receiving a reasonable percentage of fairly correct schedules is greatly augmented. The average informant is surprisingly ignorant and careless in matters of this kind and questions must be made much more simple than if they were to be placed in the hands of enumerators. A schedule of this sort should always bear a statement of its exact purpose and the person or authorities responsible for the inquiry. Otherwise, suspicion and prejudice will unite with the natural inertia of the informant and replies will not be forthcoming. Questions for this variety of schedules should usually deal with present facts, only, for it is practically hopeless to get records

of the past which are accurate enough to warrant the trouble involved.

The main advantage of this method is that a large territory may be covered at only a small fraction of the expense necessary to pay for sending out enumerators. If a reasonable number of well-filled schedules are received, they constitute good samples which are generally representative of the results as a whole and hence the work may be completed with tolerable accuracy.

This plan is extensively used by private individuals and also for government reports. Statistics of wages, unemployment, local expenditures, weather reports, etc., are regularly obtained in this manner. Where there is a law requiring the filling of the schedule and a legal penalty attached for neglecting this duty, very satisfactory results are often forthcoming. Even voluntary reports, as of the weather, frequently furnish quite regular and reliable data. This is much more often true in regard to reports sent in by picked observers at regular intervals than of those sent in only once. Most of the rules for schedules and questions given under the next title are also applicable to the foregoing. We shall discuss these under the head

Sec. 32. Schedules in Charge of Enumerators.

This is the plan followed in the leading governmental investigations and is usually too expensive to be undertaken by private initiative. It is unquestionably the best plan for most kinds of extensive inquiries.

TABLE I.

OCCUPATION AND WAGE SCHEDULE.

Name of Person Employed.	Employment.					Wages.		Unemployment.		
	Present Employer.		Nature of Work.					Number of Weeks Unemployed During Past Year.		
	Name of Person, Firm or Corporation.	Location of Plant in which Employee is Now Working.	General Name of Industry in which Now Employed.	Branch of said Industry in which Employed.	Actual Occupation or Duties of Person Employed.	Wage per Hour paid in Present Occupation (State in Cents).	Average No. of Hours Worked per Week when Employed.	Because of Lack of Work, Strike or Lockout.	On Vacation.	Total.
Busse, John	Burlington R. R.	Ottumwa, Iowa	Railroad	Car-factory	Carpenter	30	51		2	2
Yates, Richard	Dain Mfg. Co.	So. Ottumwa Iowa	Implement Mfg.	Scraper-factory	Foreman	50	60		3	3
Booth, Wm.	Clark Drug Co.	Ottumwa, Iowa	Mercantile	Drug-store	Prescription clerk	42	70		1	1
Jones, Calvin	Bates Milling Co.	Ottumwa, Iowa	Mfg. flour	Grain Elevator	Laborer	20	60	8		8
Carter, Reuben	1st Nat. Bank	Ottumwa, Iowa	Banking		Asst. Cashier	66	8		3	3
Barnes, Robt.		Ottumwa, Iowa	Draying		Drayman				1	1
Sullivan, Thos.	Central Hdw. Co.	Ottomwa, Iowa	Building trade		Bricklayer	50	8	12	1	13

In this kind of an inquiry, the schedules may be much more complete than in case of those sent directly to voluntary informants and hence the scope of the inquiry may be greatly enlarged. However, care should be taken that the schedules should be of convenient size and form for the enumerators to handle, not unwieldly folded sheets of large size which are hard to manipulate and are easily torn. The schedules should also be spaced and ruled in such a manner as to enable the eye to easily follow the line or column across the paper. Headings and subheadings should be placed in proper relationship to each other and the type used be such as to bring out the distinctions properly. Every heading, form and title should be so lucid that any person of ordinary intelligence can fully comprehend its significance. Each word and phrase should therefore be carefully scrutinized for possible double meanings or debatable interpretations. Care should also be taken to indicate in the headlines the exact degree of accuracy to which each numerical result is to be read. These simple precautions will prevent much needless confusion and loss of time to the enumerators as well as many unnecessary errors.

A sample schedule of occupation and wages is given opposite, this form being intended for cases in which the information is to be gathered from the wage earners and not from the employers. Of course, full instructions for the interpretation and use of the schedule as

well as a sample schedule properly filled out should be part of each enumerator's equipment.

Another type of schedule very popular at present is the individual card. In this case, the information concerning each separate individual is placed on a card by itself. This plan is especially desirable in case that it is intended to later group the different items in the schedule in different orders. For example, if, in the appended schedule, one desired first to classify the workers according to occupations and later according to wages and unemployment the grouping process would be greatly facilitated by the use of cards. The card system is almost the only feasible one where the record is to be continuous and constantly expanding. On the other hand, cards are much more bulky, require larger space for filing and are more inconvenient to total. For the purposes of the U. S. Census and other extensive investigations, a combination of the two methods is used. The data are first entered on schedules by the enumerators, and later punched on properly prepared cards on which titles are represented by numbers, and then tabulated by intricate electrically-operated machines. The method to be adopted is, of course, dependent on the specific characteristics of the investigation in question.

Sec. 33. The Choice of Questions.

In selecting the questions to be asked of the informant, one must differentiate between the cases in which

the filling of the schedules is left to the convenience of the informants and those cases in which the furnishing of the information is required by law. In the first case, questions must be **very simple, few in number,** and **easy to answer,** otherwise, one may feel sure that a large percentage of them will go unanswered. In the latter case, the number of questions may be considerably increased but their character of simplicity must be preserved in order to obtain answers whose accuracy is sufficient to be of value.

When enumerators are sent out, especially if they have legal powers to require answers to their questions, the questions may be more complex and more numerous. They must never be so difficult that the enumerator cannot correctly interpret them by aid of his printed instructions. If, however, the enumerator thoroughly comprehends the question he may, by several related inquiries, obtain the required data even where the meaning of the original question is not entirely clear to the informant.

The investigator must always bear in mind that each additional question means additional expense and extra work in tabulation. In an extensive investigation like the National Census, a single inquiry entails a cost of many thousands of dollars. Under these circumstances, the number of questions is strictly limited by the funds available. It becomes, then, a question of eliminating the least essential questions and retaining those deemed most indispensable.

In order to render it possible to tabulate the results, **it is necessary that the questions shall be such as may be answered by yes or no or a simple number.** If a question asks for the education of the informant an infinite number of over-lapping answers will result such as "good," "considerable," "well-posted," "high-school," etc. Such answers can scarcely be handled statistically. If, on the other hand, the number of months attendance at school is asked for, the numerical reply will be susceptible of tabulation and will approximate the result sought

In draughting a set of questions, one must, as far as possible, **avoid those which are likely to arouse the resentment of the informants or those whose answers are likely to be affected by prejudice.** If hostility is once aroused, it is difficult to get any further correct information. Questions concerning indulgence in stimulants, physical or mental infirmities, and the like, are good examples of those which provoke antagonism. The ages of women are likely to be understated in a large fraction of all answers given. When it is necessary to ask inquisitorial questions it is a valuable help to check the answer by some corroboratory question as, for example, to inquire the age in years and, at some later point in the inquiry, get the date of birth. A skillful enumerator should be able to thus unravel the truth in many cases.

Care should always be taken to see that the questions

exactly cover the points desired in the study and are not subject to double interpretation. One of the main difficulties in using results of other investigations is that the questions asked therein were not usually intended to cover exactly a similar point and a slight difference in the wording may bring vastly different results. If, for instance, one were computing the number of train-miles travelled on a certain railroad system, it would be a matter of decided importance as to whether the query was so worded as to include or exclude the distances covered by switch-engines.

To summarize, then, the questions chosen should be[1]

1. Comparatively few in number.
2. Require an answer of a number or yes or no.
3. Simple enough to be readily understood.
4. Such as will be answered without bias.
5. Not unnecessarily inquisitorial.
6. As far as possible corroboratory.
7. Such as directly and unmistakably cover the point of information desired.

Sec. 34. **Defining the Field.**

Having settled upon the problem, method of inquiry, schedules, questions, etc., the investigator must now decide upon the scope which the inquiry is to have. Both time and space being limitless, certain definite confines must be fixed beyond which the study shall not extend. If a study of incomes is to be made, it

[1] See Bowley, A L., *Elements of Statistics*, pp. 18–25.

may deal with one city, or several cities, one county or several counties, one state or the whole nation. One may compare the incomes in the different localities at the present time or in the same locality at different times.

Sec. 35. Representative Data.

Private investigators, being usually unable to cover thoroughly as large a field as desired, quite generally resort to the method of securing sample data. If it is wished to learn something of workingmen's budgets, no effort is made to obtain the record for all the families of any one community but sample families are taken which are supposed to represent the entire field. The results thus obtained are likely to be quite satisfactory if the instances are numerous and the sampling has been properly done. The study of Professor Chapin of the condition of the working people of New York City gives results in many respects quite in accord with the far more extensive investigations conducted by the Bureau of Labor, though the number of families that he studied was necessarily very much smaller than that included within the bounds of the government inquiry. There is, however, always danger of incorrect sampling owing either to accident or to conscious or unconscious manipulation on the part of the investigator in order to obtain the results desired. A Marxian socialist, desiring to prove that conditions were growing worse, would be apt to select too large a percentage of the poorest families;

an optimist would probably take too many instances from the more fortunate classes. Proper sampling may be secured simply by taking a very large number of instances at random but, when the number forms but a small fraction of the aggregate, it is better to divide the entire group to be studied into classes, ascertain as closely as possible the total number in each class, and then select samples from the various classes in the ratio of their respective numbers.

Still another modification of this plan, which is scientifically accurate but practically more difficult to apply, is to arrange the items as nearly as possible in order according to size and select samples at approximately equal intervals throughout the series. This finds application principally in the field of biology where where it is used to avoid dealing with too large a number of items.

Sec. 36. **Selection of Enumerators.**

It is almost unnecessary to remark that the quality of an investigation will depend largely on the character of the enumerators employed. Intelligence is necessary in order that the vague replies of the informants may be eliminated and put in shape for recording. But intellectual capacity is far from being the sole requisite. Diligence and integrity are just as necessary. The unscrupulous enumerator will save much effort by filling in schedules with fictitious quantities and so vitiate the entire result. Those persons directly interested in the

outcome are likely, of course, to allow their personal bias to enter into their records to a greater or lesser degree. In addition, the enumerator should be courteous and tactful in order that the work may proceed smoothly and the correct replies be elicited wherever possible.

REFERENCES.

MEITZEN, AUGUST. *Statistics*, pp. 120–129, 155–185.

BOWLEY, A. L. *Elements of Statistics*, Chaps. II and III.

YULE, G. U. *Introduction to Statistics*, Chaps. XIII and XIV.

BLOCK, MAURICE. *Traité de Statistique*, Chaps. IX and X.

BERTILLON, JACQUES. *Cours Élémentaire de Statistique*, Chaps. IV and V.

BOWLEY, A. L. *Elementary Manual of Statistics*, Chap. VIII.

CHAPTER VII.

THE COLLECTION OF MATERIAL.

Sec. 37. The Secondary Method.

When the investigation is to be of a secondary type, it is necessary to exercise considerable care in several respects, before making use of the figures gathered by others. The first essential is to know something of the reliability of the original compiler of the data and his ability to get at the facts. They may represent mere guesses and in that case, it is folly to use them as a basis for scientific work. If one is satisfied that the tables have some real claim to merit he should next proceed to determine the following facts as completely as possible.

1. From what sources the figures have been derived.
2. The definitions of the units, including the instructions to the enumerators.
3. The purpose for which the data were originally collected.
4. The methods used in collecting the same.
5. The degree of accuracy of the figures.

These points having been satisfactorily settled, the investigator is now able to make use of the figures in an intelligent manner.

Frequently, considerable discrepancies will be found in the figures in different parts of the same report.

These are often due to the omission of certain parts from one total which have been included in another similar one. Often a little careful reasoning and observation will locate the source of the error.

Estimates or figures for the same amount taken from different sources often differ widely. In this case it usually requires much greater effort to reconcile the diverging numbers. If both are from apparently reliable sources, it may be worth the effort, otherwise all but the best authenticated must be rejected or else the estimate be only made accurate to the furthest digit in which the different numbers coincide.

Sec. 38. **The Primary Method.**

After the schedules have been returned by the informant or enumerators, as the case may be, there is still, as a rule, much work to be done before the results are ready for tabulation. Each schedule must be checked over for errors and omissions. Sometimes the latter may be supplied by a second inquiry but this entails a very considerable amount of extra labor and expense and, ofttimes, surprisingly little additional information is elicited, most of the original blanks being due not to oversights but to some difficulty in answering the question.

If the figures are manifestly erroneous, they must be either corrected or rejected. Frequently an entire schedule will be found so incomplete or so badly tangled that it must likewise be thrown out. It is better, by

far, to have a smaller number of correct samples than to have a large number of incorrect ones. In the first case, the error can often be mathematically corrected with approximate accuracy; in the latter case, there is no remedy.

REFERENCES.

MEITZEN, AUGUST. *Statistics*, pp. 155–206.

BOWLEY, A. L. *Elements of Statistics*, Chap. III.

BERTILLON, JACQUES. *Cours Elementaire de Statistique*, Part IV.

BLOCK, MAURICE. *Traité de Statistique*, Chaps. IX, XI. and XII.

CHAPTER VIII.

APPROXIMATION AND ACCURACY.

Sec. 39. **Perfect Accuracy Rarely Attainable.**

As was mentioned in the first chapter, statistics as a science deals with estimates rather than with exact enumerations. If we wish to measure the total product of the coal mines of the United States, it is self-evident that the result can only be approximated. Not a single carload can be weighed with exactitude. There are likely to be errors in the number of car-loads reported from the various mines. In addition, a large number of small mines are sure to be omitted from the list. Taking all these facts into consideration, one readily sees that the total may be in error many thousands or perhaps even several millions of tons. In dealing with questions of error, however, relative and not absolute accuracy is the standard in mind. The production of coal for the United States in 1909 was over 397 million tons, hence, an error of one million tons only amounts to about one fourth of one per cent. For many purposes, an error of even four or five per cent. might not seriously vitiate the result.

Absolute accuracy is not possible in any case in which measurement is involved. By means of an ordinary ruler, one might measure the length of a needle in whole millimeters. By substituting a simple vernier,

the accuracy could readily be increased to tenths of a millimeter and, by still more refined processes, the error might be reduced to the thousandth part of a millimeter but still the measurement would never absolutely accord with the length of the needle. More refined methods simply give a constantly closer approach to exactitude, but never attain it.

Sec. 40. **Standard of Accuracy.**

While, in the physical sciences, very great accuracy of measurements is practicable, this is far from being true in the case of social phenomena. In this field a multitude of sources of error are ever present, many of which can be eliminated by no degree of care. Fortunately for the statistician, small errors are often negligible and in no way obstruct the solution of the given problem. Attempts to obtain the greatest possible degree of accuracy are, frequently, merely wastes of time. It might be possible to measure the customs revenue of the United States to the nearest cent but, for ordinary purposes of statistical comparison, such accuracy is not only superfluous but positively confusing to the mind inasmuch as the addition of extra figures diverts the attention of the mind from the fundamental digits.

For every statistical problem, there should be determined, in advance, a definite standard of accuracy for each item and every endeavor should be made to bring each recorded instance up to this standard but this

standard by no means needs to correspond to the highest degree of accuracy attainable. In taking an age census, it would probably be possible to determine the age of most persons to the nearest day but there would be no advantage in so doing. Hence, the only desideratum is to obtain data sufficiently accurate for the purposes for which they are intended or are likely to be used.

Thus, in the case of a large lake, the area in square miles would be accurate enough for all purposes but, in the case of a small reservoir, measurements in cubic feet or gallons would be appropriate. The nearest mile sufficiently approximates the length of a railroad but light waves must be measured to millionths of a millimeter. To repeat, **relative and not absolute accuracy is the desideratum.**

Sec. 41. **Round Numbers.**

Since numerous digits are confusing to the mind, it is frequently best to express quantities in round numbers even where the exact figures are available. For example, if one wishes to express the comparative populations of the United States and China to the ordinary audience, it is far better to state the population of the United States as ninety millions and that of China as four hundred millions than to give the census figures for each nation, for the hearers, in trying to sense the digits, fail to comprehend the main point which the lecturer is trying to convey.

This same use of round numbers is allowable in popular books and magazine articles and for some purposes in scientific works. If one wishes to show the comparative amount of steel produced in the United States during the last thirty years, it is perfectly correct to tabulate it as follows even if accuracy to a further degree is possible.

Year.	Steel Produced in Hundreds of Thousands of Tons.
1880	12
1885	17
1890	43
1895	61
1900	102
1905	200

The remaining figures should in no case be given if they are statistically inaccurate. If, however, they are accurate it is often advisable to state the complete numbers in a table so that someone else may use the figures in other combinations where greater accuracy may be desirable. This being done, round numbers may be used in comparisons in the text and the principal points brought, in this manner, to the reader's attention. This is probably the method most generally applicable in scholarly works.

Sec. 42. **Possible Accuracy.**

While it is very easy to determine the desirable standard of accuracy, it is by no means possible to always bring every item within this limit. The engineer

on a geodetic survey may desire to read his angles to tenths of a second but his instrument may be such that seconds only may be correctly ascertained. The collector of wage statistics probably desires to know the yearly wage of each individual to the nearest dollar but, in cases of irregular employment, the chances of getting nearer to the correct sum than ten or twenty dollars are slight indeed. The geologist would fain measure the date of the beginning of the last glacial recession to the nearest century but he must be content to approximate it in tens of thousands of years. The legislative commission is anxious to determine the exact amount of highway expenditures of the state, but many of the reports sent in are so incomplete and confused that the closest estimate must probably differ by many thousands of dollars from the correct figure. In each of these instances, the standard of accuracy is not set by the statistician but by circumstances over which he has little or no control. **In such cases, he should be careful that his report shows accuracy only to the point actually attainable and not to the point desired or to the degree indicated by the most exact of his data.**

Sec. 43. **Accuracy in Entering and Reading Figures.** The accuracy to which figures are read or are correct should, in tabulation, be stated in the heading of the column or in a footnote. All inaccurate figures except the first digit beyond the margin of accuracy should be dropped.

It is frequently better to carry the digits one place further than absolute accuracy justifies for the first inaccurate digit often represents an estimate which is somewhere near the correct quantity. In such cases, its elimination increases the error of the final result. For instance, if the length of a leaf is recorded as 2.96 cm., though it was only possible to read accurately to the nearest tenth of a centimeter, it is better to retain the final digit 6, since 2.96 is likely to be a closer approximation to the real length than 3.0 which would be the reading if this digit were dropped. **If the last correct figure is a cipher, it must invariably be entered the same as any other figure.** Thus, if, in measuring the length of a leaf, one is reading correctly to millimeters but expressing the result in centimeters, and the leaf happens to be as nearly seven centimeters long as can be measured, it must be expressed as 7.0 cm., not merely as 7 cm. The latter figure indicates that the reading is only accurate to centimeters. In other words, any leaf between 6.5 and 7.5 cm. in length would in that case be entered at 7 cm. while, if the accuracy of reading is to millimeters, and the entry is 7.0 cm. it means that the length is between 6.95 and 7.05 cm. An entry of 7.00 cm. would similarly show that the reading was accurate to hundredths of a centimeter and that the leaf length was between 6.995 and 7.005 cm.

When, for any reason, it becomes necessary to drop certain digits of a number in order to bring all items

to a uniform standard or because certain digits are beyond the limit of accuracy, one should always be careful to see that the remaining digits are correct. For instance, if it is desired to reduce the following to a uniform standard of correctness of one decimal place, the results would be as follows:

Original Number.	Correct to One Decimal Place.
27.25001	27.3
27.249987	27.2
18.20995	18.2
18.9478	18.9
18.95172	19.0
19.09162	19.1
24.05002	24.1
23.04997	23.0

All fractions over half are, in every instance, counted as whole numbers and all under half are discarded. Those exactly equalling one half may be retained or dropped at discretion.

Sec. 44. **Possible Accuracy as a Result of Various Mathematical Operations.**

An exceedingly common error is to give to figures a large degree of fictitious accuracy which arises simply from some mathematical operation. Take the following example: John is seven years old, Harry nine and George is six. Find the average age of the boys. The student is likely to proceed in this fashion.

$$7 + 9 + 6 = 22,$$
$$22 \div 3 = 7.333333333 \text{ yrs. old.}$$

Evidently, the number of decimal places is limited only by the student's industry or his sheet of paper. The answer will, oftentimes, be accepted as correct yet a moment's reflection must demonstrate its absurdity. John's age is only stated in even years. If given correctly, he may lack five months of being seven years of age or he may be seven years, five months and twenty-nine days old. The same is true of the ages of the other two boys. Not one of the items, then, is given with greater accuracy than to the nearest year and the average could not possibly attain great exactness, yet, our answer purports to state the average age to the billionth part of a year. **One must guard against such fictitious accuracy whenever numbers containing decimals, or giving a decimal as the result, are multiplied, divided, raised to a power, or the root extracted.** The following discussion may be helpful in determining the accuracy of results obtained through mathematical operations.

Accuracy in Multiplication.

If m = the multiplier.

n = the multiplicand.

x = the possible error of the multiplier.

y = the possible error of the multiplicand.

Then

$$(m + x)(n + y) = mn + my + nx + xy,$$

$$(m - x)(n - y) = mn - my - nx + xy.$$

The product evidently then is equal to

$$mn + xy \pm (my + nx)$$

But xy is so small compared to the quantity $my + nx$ that it may ordinarily be neglected and the product be considered simply as $mn \pm (my + nx)$.

The error therefore $(my + nx)$ can be readily calculated and the accuracy of the product, and the number of correct digits therein determined.

Example (accurate digits italicized):

726 × *10,2*00.

Possible error of first factor is 0.5 $= x$,
Possible error of second factor is 50 $= y$.

$$m = 726.$$
$$n = 10,200.$$

The product, however, equals $mn + xy \pm (my + nx)$. Substituting

$$7,405,200 + 25 \pm (36,300 + 5,100) = 7,405,225 \pm 41,400.$$

If xy is neglected as it may well be owing to its small size, we get as the product 7,405,200 ± 41,400 which equals 7,446,600 or 7,363,800.

The greatest absolute accuracy is therefore *7,4*00,000.

Accuracy in Division.

If a = the dividend.
d = the divisor.
x = the possible error of the dividend.
y = the possible error of the divisor.

The quotient evidently lies between

$$\frac{a + x}{d - y} \text{ and } \frac{a - x}{d + y}.$$

But

$$\frac{a + x}{d - y} - \frac{a - x}{d + y} = \frac{(a+x)(d+y) - (a-x)(d-y)}{d^2 - y^2} = \frac{2dx + 2ay}{d^2 - y^2}.$$

The possible error is evidently nearly equal to half of this quantity or $\frac{dx + ay}{d^2 - y^2}$.

Therefore, the quotient approximately equals $\frac{a}{d} \pm \frac{dx + ay}{d^2 - y^2}$.

By determining the possible error, the number of absolutely accurate digits may be determined.

Example (accurate digits italicized):

$$1{,}440 \div .012$$

In this case,

$$a = 1440$$
$$d = .012$$
$$x = 5$$
$$y = .0005$$

The quotient equals

$$\frac{a}{d} \pm \frac{dx + ay}{d^2 - y^2} = \frac{1{,}440}{.012} \pm \frac{.012 \times 5 + 1{,}440 \times .0005}{.000144 - .000{,}000{,}25}$$

$$= 120{,}000 \pm \frac{.78}{.00014375} = 120{,}000 \pm 5{,}426 +$$

$$= 125{,}426 + \text{ or } 114{,}574 - .$$

Therefore, the result is strictly accurate only to hundreds of thousands but with a strong probability of accuracy to tens of thousands place since the possible error is 5,426 + and the probable error much less.

Accuracy in Square-root.

If n = the number whose root is to be extracted,

e = the possible error of the number.

Then the correct root evidently lies between $\sqrt{n + e}$ and $\sqrt{n - e}$ and the possible error is approximately $\sqrt{n} - \sqrt{n - e}$.

This quantity is but a fraction of $\sqrt{e}$ hence the error is greatly reduced by extracting the root.

Illustration (accurate figures italicized): $\sqrt{14{,}400}$ lies between $\sqrt{14{,}450}$ and $\sqrt{14{,}350}$. The possible error is $\sqrt{14{,}400} - \sqrt{14{,}350}$. This equals 120.0 − 119.79 + = 0.21 −. But the possible error of the original number equals 50 and the square root of 50 is 7 +.

Hence, the possible error of the square root is much less than the square root of the possible error of the original number.

Accuracy of a Square.

If n = the number to be squared,
e = the possible error.

Then, the correct square must be between $(n + e)^2$ and $(n - e)^2$.

But $(n + e)^2 = n^2 + e^2 + 2ne$, and $(n - e)^2 = n^2 + e^2 - 2ne$.

Evidently, the square of the quantity then equals $n^2 + e^2 \pm 2ne$. For most purposes, e^2 is so small as to be negligible and the result may be stated approximately as $n^2 \pm 2ne$ and the number of correct digits be thus readily ascertained.

Example (accurate digits italicized):

$$(\mathit{1{,}2}00)^2 = (1{,}200)^2 \pm 2 \times 1{,}200 \times 50 = 1{,}440{,}000 \pm 120{,}000.$$

Therefore, the possible limits of the square are approximately 1,560,000 and 1,320,000, the result being accurate only to millions place.

It should be noted in each of the above cases that the possible error by no means corresponds with the probable error. The chances that the error will be the greatest possible are comparatively slight,[1] hence, as stated in Sec. 43, in entering products, quotients, or roots, one more digit should always be added than could be done if absolute accuracy to the last digit were required. By entering an extra digit, the probable error of the result is greatly diminished.

[1] For a discussion of the theory of error see Bowley, A. L., *Elements of Statistics*, page 269 f.

Sec. 45. **Compensating vs. Cumulative Errors.**

The accuracy of the final results depends, very largely, on whether the errors involved are of the compensating or cumulative type. If different people were to estimate the length of a given line, the chances would be that as many would estimate it too long as too short. The errors in measuring a line made by a pair of chainmen because of stretching the chain too tight or not taking up the slack sufficiently would tend, in the long run, to offset each other. The estimates of a thousand observers as to crop conditions compared to the previous year, while in no case accurate, would, taken together, tend to quite closely approximate the correct result. These are simply concrete applications of the law of statistical regularity.

On the other hand, if the chain used by the above-mentioned surveyors was too short, the longer the line measured, the greater the error would become. If certain reports of expenditures are missing, the large number of items present will in no way tend to offset those omitted. If women are prone to state their ages too low, the matter will not be remedied because millions of the sex are counted. The logic of trying to correct cumulative errors by mass of data is illustrated by the pun of the wag who remarked that a certain restaurant keeper was losing a little money on each meal but made it up because he had so many patrons. We may say then, in conclusion, that **when the number of items**

is large, compensating errors, if relatively small, are negligible but, on the other hand, cumulative errors always seriously affect the accuracy of the total or the average.

Sec. 46. **Accuracy of Totals.**

The strength of a chain is determined by its weakest link. Similarly, **the total can be no more accurate than its most faulty item.** No amount of compensation can overcome this fact. If a hundred railroad companies report their expenditures to the nearest cent and a single line gives its figures only to the nearest thousand dollars, it is impossible to state the total expenditure for railroad companies closer than in thousands of dollars. If this one company reports that it spends $2,673,000, we only know that the amount spent was somewhere between $2,672,500 and $2,673,500. Any amount between these figures would be correct according to the statement of the company. If the sum of the figures reported by the other companies were $16,295,472.16, we cannot legitimately add thereto $2,673,000 and then state the total in dollars and cents. We known that this answer is likely to be $500 in error either way. Hence, the correct form of the operation would be as follows:

$16,295,472.16	
2,673,000	approximately.
$18,968,000	correct approximation.

However, in obtaining the sum of several amounts one must not make the opposite mistake of dropping all digits beyond the point of accuracy of the least accurate item. For example, in adding the following column, the correct result is not 46, as one might suppose at first thought, but should be stated as below.

```
 6.321
 2.4926
21.4632
 8.
 7.3875
 2.426
-------
48.0903 summation.

48      correct approximation.
```

To have omitted the decimals, would have introduced an unnecessary error of two whole units into the total. Therefore, **all correct figures in the separate items should be retained and the rejection of the inaccurate digits be made only in the total.**

Sec. 47. **Accuracy of Averages.**

We have seen that the **absolute** accuracy of a total can be no greater than that of the most inaccurate item composing it. We shall now consider the absolute accuracy of an arithmetic average

If m_1, m_2, m_3, etc., are the estimated quantities, n in number, their respective errors being e_1, e_2, e_3, etc., then the estimated average is $\Sigma m/n$, but the largest

possible average is

$$\frac{(m_1 + e_1) + (m_2 + e_2) + (m_3 + e_3) \cdots + (m_n + e_n)}{n}$$

or $\frac{\Sigma m + \Sigma e}{n}$ or $\frac{\Sigma m}{n} + \frac{\Sigma e}{n}$.

Therefore, the average may be correctly written $\Sigma m/n \pm \Sigma e/n$. But $\Sigma e/n$ would be the average possible error of a single item. **Hence, the possible error of an arithmetic average is equal to the average possible error of the items in the series.**

In obtaining this possible error, we assumed that all the errors were in a similar direction. As a matter of fact, however, the chances that this will be true are very remote when the number of items is large and the errors are of the compensating type. The **probable** error of the arithmetic average is therefore only a fraction of the **possible** error. If E = the possible error of the arithmetic average, the probable error of the same is approximately $E/\sqrt{n}$.[1]

This fact that the probable error of the average of a number of items is less than that of any single item is of great value in scientific work. The physicist makes a large number of observations of the same phenomenon and averages the results. The surveyor repeats the measurement many times in determining an angle accurately. The sociologist obtains observations in many different localities in order that

[1] For proof see Bowley, *Elements of Statistics*, pp. 303–315.

peculiar conditions in one place may be offset by reverse surroundings in another. Even the personal bias of the observers tends to be eliminated by the averaging process.

Still, one must not go to the opposite extreme and conclude that an average always reduces the error to a negligible quantity. Biased errors are not mutually corrective and hence are in no way reduced in the average and even unbiased errors may be large enough to greatly vitiate the average. When the items are few in number, the effect of errors is especially serious.

Sec. 48. **Locating the Decimal Point.**

In performing mathematical operations with a slide rule, one obtains merely a sequence of figures and the beginner is often puzzled as to the correct location of the decimal point in the result. One should learn to determine this correctly by inspection but, until this is accomplished, it may be properly placed by means of the following empirical rules.

Multiplication.

1. Consider the first significant digit in the multiplier as a unit and the remaining figures as decimals of the same. Do the same for the multiplicand.

2. Obtain a mental product of these two numbers and note whether it contains one or two integral digits.

3. Add to the number of integral digits thus obtained the total number of integral digits not used in the pre-

liminary multiplication contained in the multiplier and multiplicand taken together and, in case either the multiplier, multiplicand, or both should be wholly decimal, subtract from the sum thus obtained the sum of all ciphers preceding the significant digits in the multiplier and multiplicand and the decimal points in each entirely fractional factor. The result, if positive, indicates the number of integral digits in the product; if negative, the number of ciphers preceding the first significant digit of the product.

Example I.

42,000 × .025 gives 105 as the sequence of figures.

$$4.2 \times 2.5 = 10 + \text{ or two digits.}$$

The multiplicand contains four other integral digits. The multiplier contains a decimal point and one cipher.

$$2 + 4 - 2 = 4, \text{ the number of integral digits.}$$

The product, therefore, equals 1,050.

Example II.

.036 × .024 gives 864 as the sequence of figures.

$$3.6 \times 2.4 = 8 + \text{ or one digit.}$$

The factors contain two decimal points and two initial ciphers or four digits.

$1 - 4 = -3$ or the number of initial ciphers in the product.

Hence, the product equals .000864.

Division.

In division, as well as in multiplication, it is usually possible to locate the decimal point in the result by inspection, but, in case there is difficulty in so doing, the following rules and table may prove of assistance

Considering digits to the right of the decimal point as negative and to the left as positive, count from the decimal point to the first significant figure in the dividend and record the result. Repeat for the divisor. From the number determined for the dividend, subtract, algebraically, the number found for the divisor and add to the remainder the number found under the proper headings in the table below.

Let D = first significant figure of dividend.
d = first significant figure of divisor.

TABLE II.

DETERMINATION OF DECIMAL POINT IN DIVISION.

Characteristics of Dividend and Divisor.		Dividend Equal to or Larger than Divisor.		Dividend Smaller than Divisor.	
		$D = d$ or $D > d$	$D < d$	$D = d$ or $D > d$	$D < d$
Dividend unity or greater.	Divisor unity or greater	+1	0	0	−1
	Divisor wholly a decimal	0	−1		
Dividend wholly a decimal.	Divisor unity or greater			0	−1
	Divisor wholly a decimal	+1	0	0	−1

Examples:

1. $.002 \div .04$

In this case we obtain for the number of the first significant

digit in the dividend -3 and for the number of the first significant digit in the divisor -2. But $(-3)-(-2)=-1$. The dividend is a decimal, the divisor a decimal, the dividend is smaller than the divisor, and $D < d$, hence, by the table, we add (-1),

$$(-1)+(-1)=-2.$$

Therefore, the quotient $= .05$.

2. $$1{,}000 \div .025.$$

In obtaining the number of the first digit, $4-(-2)=6$. From the table, add (-1).

$$6+(-1)=5.$$

Therefore, the quotient is 40,000.

Squares.

Follow the same rules as for multiplication.

Square Roots.

Divide the given number into periods of two figures each, counting each way from the decimal point. Now count the number of periods from the decimal point to the first significant digit. This will give the number of the first significant digit of the root.

Examples:

1 $$\sqrt{36'00}.$$

Counting periods to the first significant digit gives $+2$. The root, then, is 60.

2. $$\sqrt{.00'00'81}.$$

The number of the first significant digit is -3. The root, then, is .009.

REFERENCES.

Bowley, A. L. *Elementary Manual of Statistics*, Chaps. II, III, IV.

"Student." *The Probable Error of a Mean*, Biometrika, Vol. VI, p. 1, 1908.

Bowley, A. L. *Elements of Statistics*, Chap. VIII.

PART III.

ANALYSIS OF THE MATERIAL COLLECTED.

CHAPTER IX.

TABULATION.

Sec. 49. **General Rules.**

At first thought, it would seem the simplest thing in the world to construct a table, but the beginner who attempts to tabulate a complex group of figures will quickly discover that the simplicity of the operation is far more apparent than real. In fact, when a scientific tabulation has once been made, it is often found that a large share of the work of analysis is completed. This is so far true that, until quite recent years, statisticians looked upon the table as the "ultima thule" of their efforts.

In beginning a tabulation, the first question that arises is whether the figures should be grouped in one or several tables. A single table has the merit of compactness and the data are thus brought into proximity. The table, however, if too large, becomes confusing to the eye and there is great difficulty in following the lines and columns at a glance. This difficulty may be partially obviated by varied modes of

ruling and spacing but, in general, it is better, where practicable, to break the table up into several separate sections.

Each table should be a unit. Rarely, indeed, should one attempt to demonstrate in the same table several comparisons of different natures. For each distinct purpose, there is usually one tabular form which is best suited to bring out the point desired. If another topic is included, the result is that each result is obscured. For example, it would ordinarily be unwise to attempt to show statistics of wages and unemployment in the same table for the groups which would best illustrate earning capacity might differ considerably from those which would bring out the characteristics of unemployment.

Another matter to be decided upon is whether the table shall show absolute figures, percentages, or both. This depends upon the kind of comparisons to be made. If one wishes to compare the wheat crops of various nations, it is manifestly useless to reduce the amounts to percentages of the world's crop for the percentages would be in the same ratio as the absolute figures but, if the object is to compare the amount of insanity among city and country dwellers, the actual numbers of insane in each place would tell us nothing and only when reduced to ratios or percentages does any meaning appear. In a reference work or in original investigations, the absolute figures should always accompany the

percentages so that they will be available for other studies of a different nature. Percentages or ratios, as well as totals and averages, are essential features for the majority of tables.

The number of separate headings or columns to be used is a third query which must be answered. The more minute the subdivisions, the greater the accuracy attainable. On the other hand, a multiplicity of headings prevents the proper emphasis being given to the main facts and tendencies shown by the statistics. The exact number of divisions is something that depends on the specific data in each case and must be left to the judgment of the statistician. In general, it is more satisfactory to use a few main divisions with several subheadings under each than to have a large number of coordinate titles. It is possible, by using this method, to enter total or percentage columns for each of the main divisions thus bringing out distinctly the principal ideas while still reserving the minor columns for details. If the tables are large so that these total or percentage columns fall too far apart for easy comparison, it is best to enter these together in a separate summary table so that the eye can take in the general results at a glance. In a synoptical table of this sort, it is often preferable to simply state the results in round numbers, since for the reasons mentioned in Sec. 42, a large number of digits tends to confuse the mind and prevent a proper grasp of the

meaning of the figures. The column may, for example, be entitled simply "Expenditures in tens of thousands" or "Product in millions of bu.," thus eliminating several superfluous digits. It is rarely advisable to drop the final digits in the primary tabulation and it should not be done in the summary tables if they are likely to be used primarily for reference rather than for the purpose of drawing general conclusions.

Sec. 50. **The Title of the Table.**

The title of the main table as well as of each of the subheadings should be complete and self-explanatory for the reader will seldom care to take the trouble to hunt up references in the text or footnotes in order to learn the significance of certain headings. A common error which should, of course, be avoided, is to make the title of too small scope to cover all the data in the table. Another frequent mistake, which is more serious is to have a title which is so indefinite as to permit of a double meaning, as, *e. g.*, "Percentages Engaged in Various Occupations, by Nationalities" may mean either the percentage of workers in a given occupation belonging to each nationality or, on the other hand, the percentage of the workers of each nationality engaged in the given occupation.

The column headings should, where measurements are included, invariably state the unit used, as "Height in inches," "Price in dollars," etc.

Titles should always be in Roman characters rather

than in script, the typing of each heading should correspond in size and prominence to its respective importance, and all coordinate headings should appear in type of like size and style. The second requirement prevents typewritten tables from being satisfactory if the headings are at all complex. When there are a large number of coordinate headings, some systematic sequence must be determined upon. For instance, the states of the Union might be placed in order of geographical location, area, population, alphabetically according to their names, or with some other logical criterion in mind.

Sec. 51. **Form of the Table.**

The table should always be roughly drafted in its complete form before any of the ruling of the permanent table is begun or any figures are entered. This is imperative in order that the table may be adjusted to the size of the sheet, the proper width of the columns be calculated and the correct arrangement of the headings be provided for. Space must usually be allowed for percentages or averages as well as for totals.

All numbers which are to be compared must be placed close together and, wherever possible, they should be placed in the same vertical column rather than the same horizontal line. Columns which are intended for comparison should be placed as close to each other as possible. Totals, averages, and percentages should invariably be placed adjacent to each other. Since

these are usually the parts of the table in which the general reader is primarily interested, their normal position is at the top or in the left-hand columns of the table. Custom has placed them at the close of the table instead, but the U. S. Census Bureau has followed the new policy and its example seems worthy of imitation. Unimportant totals, used only as a check, should be relegated to the customary final position.

The rulings in tables should, like the headings, indicate the importance of the various subdivisions. The principal groups should be separated by heavy or multiple-ruled lines and the breadth of the lines should decrease as the subdivisions become lower and lower in rank. In tables of any considerable size, the figures should cross the paper in horizontal bars five to eight lines in depth with narrow blank spaces between the bars.[1] In this way, the eye is saved much difficulty in locating a desired number.

Since it is often impossible to provide sufficient columns to specify all the different types of data enumerated, the odds and ends are usually placed together in a "Miscellaneous" group. This saves room and also avoids large blank spaces in the table, which are undesirable, since they confuse the eye in its effort to follow columns or horizontal lines. Exceptional items should be marked with an asterisk or number referring to an explanatory note, similarly marked, at the foot

[1] See Table VII, accompanying Sec. 62.

of the page. If the exceptions are too numerous, or if the "Miscellaneous" group is too large, the value of the table is likely to be seriously affected since the results are either rendered incomplete or lack homogeneity.

Sec. 52. **Accuracy in Tabulation.**

Care should be taken to have every item in a table accurate, for the discovery of a few errors is sure to throw doubt on the merit of the work as a whole. To obtain accuracy, a regular system of checks is necessary. In the first place, each item should be gone over to see that the original entries are correct. This having been decided in the affirmative, the rest of the operation is a mechanical process. To check the totals, it is often feasible to add the items both in vertical columns and in horizontal lines. These partial totals should then be summated and, if the same grand total is obtained in each case, the additions are almost certainly correct. In cases in which this method is inapplicable, each column should be re-added.

Percentages may be checked by adding together to see if their sum equals one hundred per cent. Averages may be multiplied by the number of items and the product compared with the total.

Multiplications and divisions should ordinarily be performed twice, preferably by two persons. By using such means as this, the chances of error may be reduced to a minimum.

Sec. 53. **Analysis of Results.**

The results disclosed by a tabulation are seldom fully revealed at a glance. Much is therefore added to the value of a table if it is accompanied by a written analysis which points out the principal conclusions which may be deduced therefrom, the possible errors involved, and the probable causes of the phenomena. The power to analyze a table, interpret the results correctly and state the conclusions lucidly and succinctly is one of the characteristics indispensable in a good statistician.

REFERENCES.

Bowley, A. L. *Elements of Statistics*, Chaps. IV and VI.
Bowley, A. L. *Elementary Manual of Statistics*, Chap. VI.
Bertillon, Jacques. *Cours Élémentaire de Statistique*, Chap. V Annexe I.

CHAPTER X.

SIMPLE DIAGRAMS.

Sec. 54. **Use of Diagrams.**

Figures, at best, are not easy things for the mind to grasp and hold long enough for purposes of comparison. When read to an audience, they become practically meaningless. In a book, their essential indications can only be ascertained by careful scrutiny. One of the chief aims of statistical science is to render the meaning of masses of figures clear and comprehensible at a glance. To attain this end, many devices have been invented to supplement or explain the table, of these, graphic illustrations being the most common. This chapter is devoted to an explanation of some of the simpler means used for this purpose.

Sec. 55. **Cartograms.**

Many phenomena which vary with geographic location are best illustrated by means of cartograms or statistical maps. Several varieties of maps may be used, each of which has its special merits. If only a single map is to be made or if printing costs are not prohibitive, various colors and shades may be used with great success. This plan is best exemplified by the "Statistical Atlas of the United States" published by the Census Bureau. Here we find a large collection of

maps, attractive in appearance and easily interpreted.

Since color printing is rather expensive, the same ends can frequently be attained at a less cost by using various modes of barring or cross-hatch work to indicate the varying degrees of density to be recorded. The rainfall maps printed in the newspapers are good examples of this class.

A third variety of cartograms, which has shown exceptional merit for certain purposes, is the dotted map. If one wishes to indicate the wheat production of the United States, he simply places a dot for every hundred thousand bushels raised in a certain section. The amount of product indicated by a single dot should be such that the dots will be placed quite close together in the regions of greatest density. Professor H. C. Taylor, of the University of Wisconsin, has used this method extensively in his studies in Agricultural Economics.

Sec. 56. Pictograms.

The baker who was trying to impress the public with the small size loaf given by his rival, Smith, advertised for some time that he sold a sixteen ounce loaf while Smith's weighed but twelve. The advertisements had little effect. When he inserted the pictogram shown in Fig. 1 the response was instantaneous. This illustrates the importance of graphic methods for illustration. A great variety of devices are used, the commonest and simplest probably being the bar dia-

gram. The size of the number is, in this method, simply represented by the length of the bar. Fre-

COMPARATIVE PICTOGRAMS

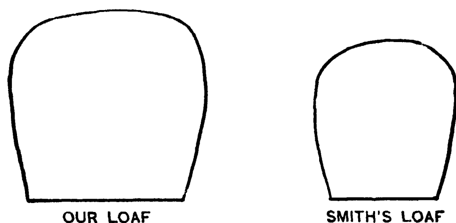


FIG. 1.

quently, the composition of the number is also represented by the shading of the bar as illustrated in Fig. 2. This has the advantage of bringing the whole study into one diagram and the disadvantage that all the sections, except the first, are rendered rather difficult of exact comparison since they do not originate or terminate on the same straight line. The plan may be modified by making a separate set of bars for each group.

For the illustration of more complex mathematical relationships, the use of bars is usually insufficient. It is necessary in such cases to resort to figures of two or three dimensions. Fig. 3 represents a comparison of the hourly wage, length of working day, distribution of expenditures of the workers, and number of workers employed in industries *A* and *B*. It may be said that, in general, as the number of factors to be compared increases, the accuracy of comparison of some of the

groups is diminished. In the illustration, while it is quite easy to compare the number of workers, the

BAR DIAGRAMS.

Population of Cities by Race and Nativity.

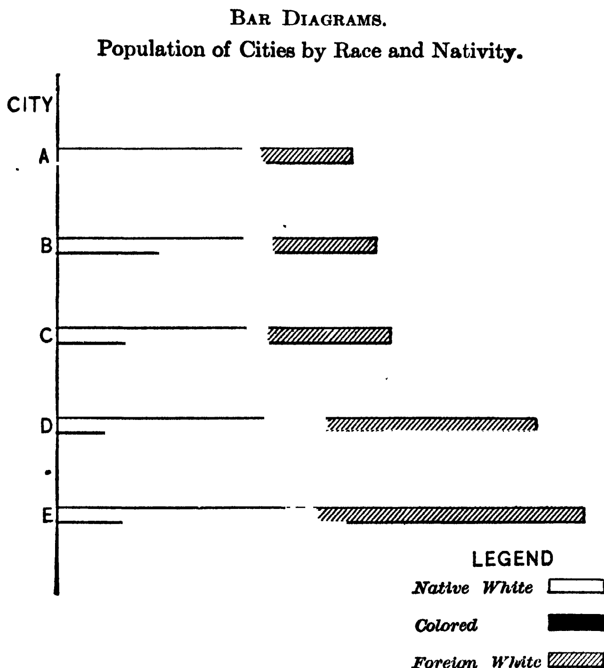


FIG. 2.

hourly wages, and the hours worked per week, both the absolute amounts and the percentage of the income spent for the different items are not easily comparable because of the varying shape of the rectangles or parallelopipeds involved.

Another useful variety of pictogram is the circle divided into sectors as shown in Fig. 4. These sectors may be colored or shaded as desired. In the illustra-

BLOCK PICTOGRAMS.

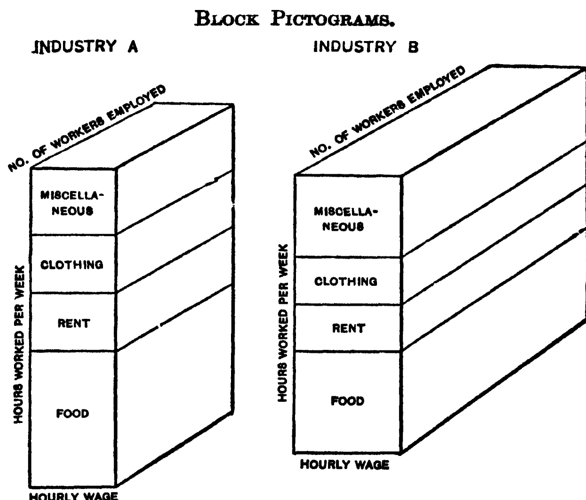


FIG. 3.

tion, the size of the circles represents the total cultivated area while the angular dimensions of the sectors show the relative importance of each crop in the states involved.

For comparing relative volumes, cubes or spheres are frequently used and, in popular works, we often see pictures representing the articles in question as, for example, a line of ships whose sizes represent the merchant marines or navies of different nations or a

row of bales showing the comparative amounts of cotton produced. It should always be remembered that, in all figures depicting area, the dimensions must

CIRCLE PICTOGRAMS.

Cultivated Area Devoted to Various Crops.

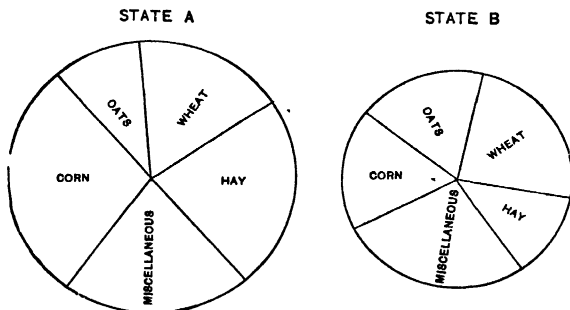


FIG. 4.

vary as the square roots of the areas represented and, if volumes are to be illustrated, the dimensions must vary as the cube roots of the contents. Failure to observe this rule often results in grotesque misrepresentations.

REFERENCES.

The Statistical Atlas of the Unites States (Census of 1900).

BAILEY, WM. BACON. *Modern Social Conditions.* Century Co., N. Y., 1906. Pp. 54–66.

BOWLEY, A. L. *Elementary Manual of Statistics*, Chap. V.

BERTILLON, JACQUES. *Cours Élémentaire de Statistique*, Chap. XI.

BLOCK, MAURICE. *Traité de Statistique*, Chap. XIII.

CHAPTER XI.

FREQUENCY TABLES AND GRAPHS.

Sec. 57. Use of Frequency Tables.

In the study of most groups of natural phenomena, we find that the different items have varying characteristics. Some elm trees are tall, others are short; some men are rich, others are poor; some cities are large, others are small. Things which thus vary in size may be spoken of as **variables.** In the cases mentioned above, the variation was between different things of the same kind or species. It is also commonly true that the same thing changes in characteristics with the passage of time. Trees grow taller, men become richer or poorer and cities become larger. This method of change, in which a period of **time** is necessarily involved, may be designated as **historical variation.** It is discussed more fully in Chap. XV.

In studying things of the same variety, the work may usually be facilitated by dividing the items into **classes.** The simplest mode of classification is to group all the instances under two headings, the determining factor being whether they do or do not possess a given characteristic. Thus, we may classify people as sane or insane, workmen as employed or idle, flowers as white or colored, men as short or tall. For some purposes, this **division by dichotomy,** or cutting in two, may be most

satisfactory but, in many cases, the difficulty arises that there is no distinct dividing line. Thus, it is impossible to say at just what point a man ceases to be short and becomes tall. It is, therefore, necessary to lay off arbitrary boundaries between the two classes. But, if classes are to be thus arbitrarily established, it is often much more advantageous to set up a larger number of them rather than two only. In practice, this is usually done by dividing the whole group into classes of equal width. Thus, if the tallest tree in a group is 39 feet and the shortest 16 feet high and it is desired to divide the entire group into five classes, the boundary lines would preferably be fixed on the round numbers 15, 20, 25, 30, 35, and 40 feet. These boundary lines are known as the **class-limits** and the distance between the two limits of any class is designated as the **class-interval.** In the instance cited above, the class-interval would evidently be 5 feet. A table formed by thus dividing a group into a number of smaller more homogeneous classes and indicating the number of items to be found in each class is known as a **frequency-table.** The number of items falling within a given class constitutes the **size** of that class or its **frequency.**

If one is told that there are one hundred men in a community and that their total possessions aggregate one million dollars or that they possess, on an average, ten thousand dollars worth of property each, he still knows very little about the welfare of the neighborhood.

Does each of the men have an equal share of the wealth or does one man own $990,100 and the remainder $100 each? In studying the distribution within the community we find use for a frequency table. The mode of procedure is to divide the population into comparatively small classes according to the amount of their wealth and then compare the relative size of the various classes or in other words the **frequency distribution.** If a somewhat similar community of a hundred men

TABLE III.

SIMPLE FREQUENCY TABLE SHOWING DISTRIBUTION OF WEALTH.
(Small Class-interval.)

Size of Items or Wealth per Man. m	Frequency or No. of Men in Class f
$ 0–$1,000	5
1,000– 2,000	8
2,000– 3,000	10
3,000– 4,000	12
4,000– 5,000	14
5,000– 6,000	10
6,000– 7,000	9
7,000– 8,000	10
8,000– 9,000	6
9,000–10,000	2
10,000–11,000	3
11,000–12,000	1
12,000–13,000	2
13,000–14,000	0
14,000–15,000	2
15,000–16,000	1
16,000–17,000	1
17,000–18,000	1
Above 18,000	3
	$n=100$

were thus arranged, we might find results something like the following providing there was a fairly equal distribution.

We observe here that the number of men per class, or the frequency, rises to a maximum (known as the mode), in the $4,000–5,000 subgroup and then gradually falls off, but with some irregularities. Let us see what the effect will be if we increase the class-interval.

TABLE IV.

SIMPLE FREQUENCY TABLE SHOWING DISTRIBUTION OF WEALTH.
(Larger Class-interval.)

Wealth per Man. m	No. of Men in Class. f
$ 0–$3,000	23
3,000– 6,000	36
6,000– 9,000	25
9,000–12,000	6
12,000–15,000	4
15,000–18,000	3
Over 18,000	3
	$n=100$

It is evident that increasing the class-interval or the distance between the class-limits from $1,000 to $3,000 has increased the regularity of the rise and fall of the figures in the second column. To gain this regularity, however, we have been compelled to sacrifice, to a certain extent, the details of the picture. Here we have, then, the same old conflict between symmetry of general outlines and accuracy of detail. The exact number of classes to be used must always be left to

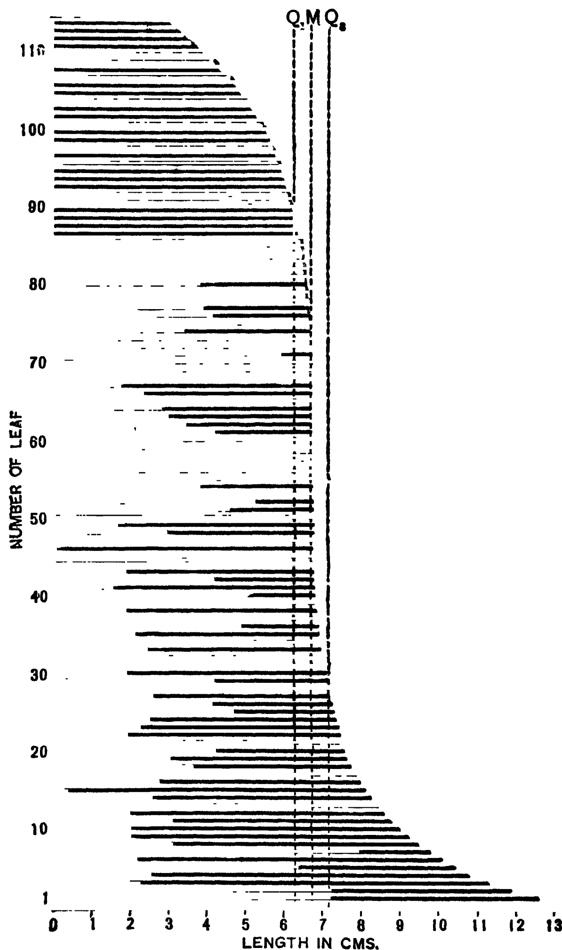

Q1 M Q3
110
100
90
80
70
60
50
40
30
20
10
1
NUMBER OF LEAF
0 1 2 3 4 5 6 7 8 9 10 11 12 13
LENGTH IN CMS.

FIG. 5.

the judgment of the statistician but, in general, the number should be as large as it may be made without sacrificing the approximate regularity of the progression.

The inquiry that naturally next suggests itself is whether the regular rise and fall noted in the preceding tables is simply an arbitrary assumption, a characteristic of the distribution of wealth, or a feature common to many varieties of phenomena. A little investigation will show the last to correspond to the facts.

If one hundred and thirteen leaves are picked, **purely at random**, from a given tree and then arranged in order of their lengths, and if a line is drawn on the paper corresponding to the length of each leaf, the results will appear as shown in Fig. 5. It will be observed that, near the extremes, the lengths change rapidly while, from the fortieth to the sixtieth leaves, the lengths are practically constant. This means that, if placed in groups having a class-interval of one centimeter, those classes between three and five and between seven and thirteen centimeters in length must contain few leaves while the class of six to seven centimeters in length would include over half of all the leaves. Evidently, then, we find the same tendency to a rise and fall in the size of classes in natural as well as in economic phenomena.

In an actual experiment of throwing three dice one hundred and ninety-six times, the following results were obtained:

TABLE V.

FREQUENCY TABLE SHOWING RESULTS OF THROWING THREE DICE.

No. of Spots (Size of Item). m	No. of Times Occurring (Frequency). f
4	1
5	4
6	11
7	10
8	24
9	22
10	22
11	32
12	17
13	23
14	9
15	7
16	7
17	4
18	3
	$n = 196$

This shows us that, in a matter of pure chance, the rise and fall of the frequencies occurs exactly the same as in the case of natural phenomena, and that the number of spots thrown fluctuates about the mode which is apparently close to eleven.

From these few examples, to which others might be added indefinitely, we can deduce the following law: **Both chance and natural phenomena tend to fluctuate about a norm known as the mode. The large majority of the items are usually grouped near the mode and, as the distance from the mode becomes greater, the items become rapidly fewer in number. In natural**

phenomena, a maximum deviation may be approximately determined beyond which no items occur. This law which, as we have seen, was discovered by

FREQUENCY LINE DIAGRAM.
Results of 196 Throws of Three Dice.

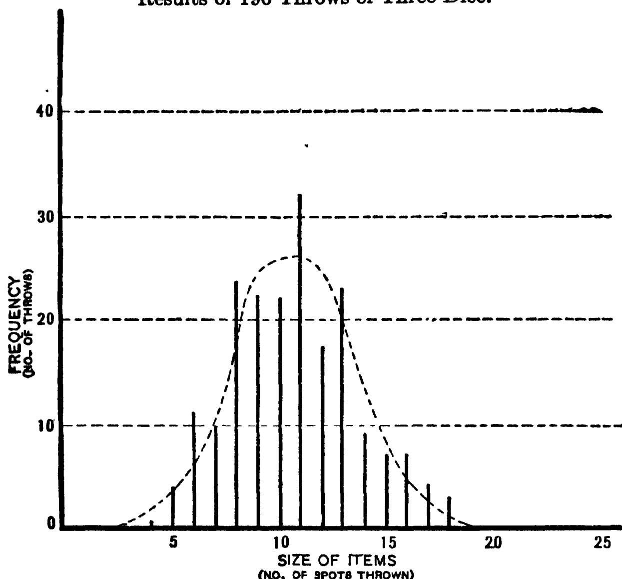


FIG. 6.

Quetelet is the basis upon which a large part of the science of statistics is built. In the leaf lengths, illustrated in Fig. 5, the mode is evidently located at about the fifty-sixth leaf length, since at this point the right-hand margin of the lines becomes most nearly

vertical, indicating the largest number of leaves within a given variation in length. We would say, therefore, that the modal length of this variety of leaves is about 6.7 cm.

Another inquiry likely to occur to the reader is whether an increase in the number of items would affect the location of the mode, that is, if five hundred leaves had been used, would it have changed the results. Experience shows that the only effect of using a larger number of items, provided the smaller number selected were fair samples, is to obtain a greater regularity in the variation in the sizes of the classes. This is due only to the fact that, the larger the number of items, the greater the chances of obtaining a fair sample of that class of objects in general.

Sec. 58. **Classification in Frequency Table.**

We have already noted the comparative merits of small and large class-intervals. Several points need also to be kept in mind. Class-intervals should all be equal, that is, the classes should be of uniform breadth and their limits should **all** be entered in the proper column whether any items occur in the class or not. If this is not done, errors in plotting the results are almost sure to be made. The size of the items of the class may be indicated by a single figure, as 3 cm. In this case, the class contains all items between 2.5 and 3.5 cm. When the class-interval is greater than one unit the titles may be entered as 3–7 cm., 8–12 cm.,

etc. The first class, then, includes all items between 2.5 and 7.5 cm. Since items, when originally measured, are usually read to the nearest unit, the foregoing is the most generally applicable system of headings.

If, however, the items are measured purposely for use in this frequency table, it is often practicable to make the classes read 3–7 cm., 7–11 cm., 11–15 cm., etc. Now, the limits of each class are on the even unit and the item 6.99 cm. in length would fall in the 3–7 cm. class.

It is always advisable, where possible, to so arrange the classes that the mid-point of each falls on an even unit and not on a fraction. This facilitates multiplication and other mathematical operations, in those cases in which the class is considered as a uniform whole. One should constantly bear in mind that, in the study of natural phenomena, the items do not fall on whole numbers and that the actual measurements are distributed more or less evenly through a group the boundaries of which are wholly artificial and arbitrary. Nature, itself, recognizes few sharp dividing lines. The classification adopted must, therefore, be purely arbitrary.

Sec. 59. **Continuous and Discrete Series.**

Since the size or weight of natural objects is likely to fall at any point whatsoever between certain limits and can never be determined with mathematical

exactitude, but is always measured by approximation, a number of such recorded measurements constitutes a **continuous** series. On the other hand, in throwing dice, it is never possible to obtain anything but an integral number of spots. A collection of items of this variety constitutes a **discrete** or broken series. A record of wages paid in a factory is likely to be a distinctly discrete series, for the wage is usually a certain number of dollars per week, the money unit seldom being smaller than half a dollar. Hence we would find no one receiving $14.39 a week and gaps would appear in the scale. In the last analysis, every series measured in a money unit must be discrete, since the smallest money unit may always be considered as divisible.

Sec. 60. **Frequency Graphs for Discrete Series.**

We have found graphs very serviceable in presenting other varieties of statistics to the eye and we shall see that they are just as useful in illustrating the frequency table. The simplest mode of illustration for a discrete series is the line or bar frequency diagram. This is especially useful in those instances in which the number of units in the scale of items is small enough so that each may be indicated by a separate line or bar. This is well illustrated in Fig. 6, which represents the frequency of various dice throws, as shown in the table in Sec. 57. Since there are no items occurring at fractional points on the scale, a

smooth curve, if drawn, must not be interpreted as indicating the existence of frequencies to be interpolated in the intervals between the integers but only as an indication of the **normal** frequencies **at each of the integral points.** The bell-shaped curve simply represents the relative height to which the vertical lines would extend provided the number of throws was infinite. This bell-shaped form, which tends to recur constantly in the study of natural phenomena, is known as the normal frequency curve or sometimes as the normal curve of error. In the given experiment, ten spots were thrown twenty-two times. By the curve, we see that, normally, both would occur the same number of times, the normal number of instances being about twenty-six. In the experiment, eighteen spots were thrown three times and three spots not at all, yet it is evident that one should have occurred as often as the other, since each dice contains a *six* and a *one* on its respective faces. The number of *sixes* thrown in this experiment was abnormally large and the number of *ones* abnormally small. **The object of smoothing, then, is to eliminate accidental variations and establish normal tendencies.**

Sec. 61. **Rectangular and Smoothed Frequency Graphs or Histograms.**

It often happens that, in either a discrete or continuous series, there is a great difference in size between the largest and smallest items and that instances occur

FREQUENCY GRAPHS SHOWING HEIGHTS OF CORNSTALKS.

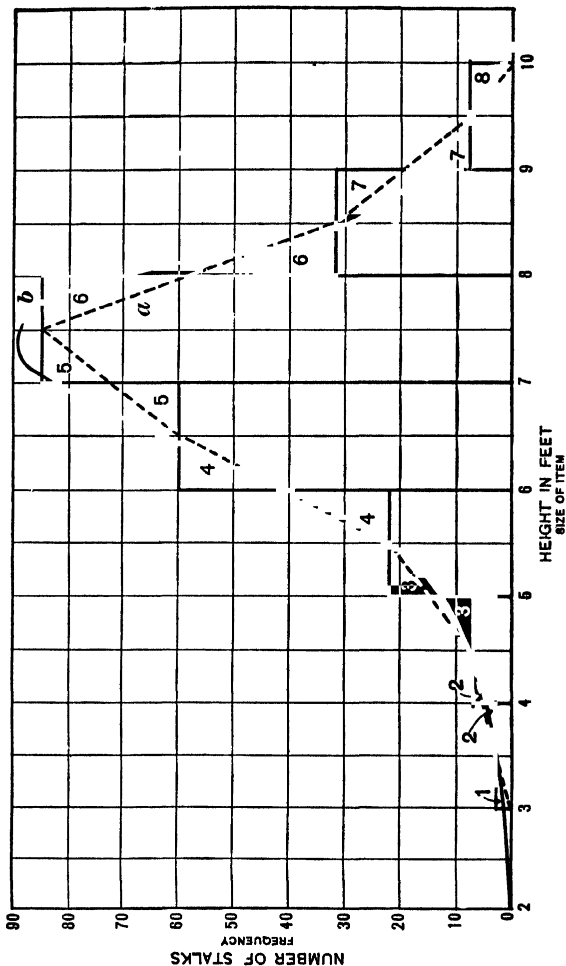


FIG. 7.

at a great number of points between the two extremes. In the case of a continuous series, as of the leaves considered in Sec. 57, an item may be any one of an infinite number of lengths. In such cases, it is clearly impracticable to place a line at each measurement at which one or more items occur. The data must be divided into classes with arbitrary dividing lines and each group is then treated as a whole. Were we to measure the height of cornstalks in a field, it would naturally be possible to find one at almost any assigned length between certain limits. If we measured a representative group of stalks, the results might be something like the following:

TABLE VI.

FREQUENCY TABLE SHOWING HEIGHTS OF CORNSTALKS.

Height in Ft. (Size of Item). m	No. of Stalks (Frequency). f
3– 4	3
4– 5	7
5– 6	22
6– 7	60
7– 8	85
8– 9	32
9–10	8
	$n=217$

This table could be illustrated by the rectangular diagram or **histogram** shown in Fig. 7. This series of rectangles illustrates fairly accurately the relative size of the various classes but, since the class boundaries are arbitrarily chosen, different sizes or arrangements

of classes would give noticeably different results. Evidently, if a larger number of items were measured and the class-intervals made narrower, the steps would decrease in size and gradually approach in form a smooth curve. The rectangular histogram is a better representative of the 217 cornstalks actually measured than any smooth curve would be, but the smooth curve would be much more representative of all the stalks in the field. Since it is the latter question in which we are usually interested, our aim is to smooth the graph so as to approach a curve which will be as typical as possible of the field as a whole.

A common method of approaching this ideal is simply to connect the outer extremities of the base of the graph with the midpoints of the tops of the rectangles as is shown in the dotted line *a* in Fig. 7. This roughly approximates the correct outline, but introduces one or two errors.

First. Though the area included in the final figure and that included in the rectangular histogram should be identical, it will be observed that when, in the figure, each included triangle is numbered to correspond with an excluded triangle, excluded triangles 1 and 8 have no corresponding included areas, hence, the line *a* encloses slightly too small an area.

Second. If the limits of the central class had been fixed at 7.4 and 7.6 instead of 7 and 8 it probably would have towered. relatively, somewhat higher than in the existing form.

Third. It is probable that stray items would occur outside the limits chosen, that is, there might be stunted stalks just under three feet or an occasional giant towering above the ten foot limit. The probabilities in this respect may best be dealt with from information outside of the table itself.

Keeping these three points in mind, we can now proceed to draw the final smooth curve or **smoothed histogram** to show what we believe to be the actual distribution of heights of the cornstalks in the field. This curve is shown in the continuous line *b*. We see that it has a marked resemblance to the bell-shaped normal curve obtained from the dice diagram.

In practice, it is very common to omit entirely the construction of the rectangular histogram and simply plot the frequencies at the midpoints of the classes, getting as a result the **frequency polygon** *a*, as the first product. This is the easiest method but it is more difficult to smooth properly and, when accuracy is desired, it is better to proceed as has been done in this case. If several frequency graphs are to be plotted on the same sheet for purposes of comparison, either frequency polygons or smoothed histograms must usually be used, since the many lines in a rectangular histogram are too confusing, and the vertical lines usually coincide.

When it is desired to smooth a frequency polygon, one should remember that it is really derived from the

rectangular histogram and proceed accordingly. This will mean that the top of the curve usually overtops the highest point of the frequency polygon, especially when the classes are rather large. All sharp turns should be avoided, the curve should change direction as little as possible, and, in most cases, irregular fluctuations should be smoothed out. The extent of the smoothing necessary or permissible must depend largely on the specific data involved. If the data consist of records of natural or chance phenomena which normally approach a symmetrical curve of error, smoothing may be freely indulged in, but if economic or sociological data are involved, considerable irregularities may really exist in the normal curve and, as a result, only the minor irregularities should be eliminated. When all the data are at hand, it must be remembered that smoothing brings out tendencies but obscures actual facts.

Since nearly every smoothed histogram representing a continuous series begins with an infinitely small number of instances and decreases again slowly to zero, it should begin and end on the base line. While no absolute rule can be laid down, the curve should, normally, reach the base about the middle of the base of the next class outside that in which the extreme instances lie.

Sec. 62. Comparative Histograms.

Histograms are very useful for comparing the structure of two or more groups of data. Table VII and

Fig. 8 give illustration of histograms showing the comparative wages in three states, *A*, *B*, and *C*. The number of wage earners in state *A* is comparatively small and the bulk of the workers receive wages between $8 and $16 per week. In state *B*, two diverse groups are distinctly noticeable. This may be a state where many women and children are employed at low

ABSOLUTE HISTOGRAMS SHOWING COMPARATIVE WAGES IN STATES *A*, *B*, AND *C*.

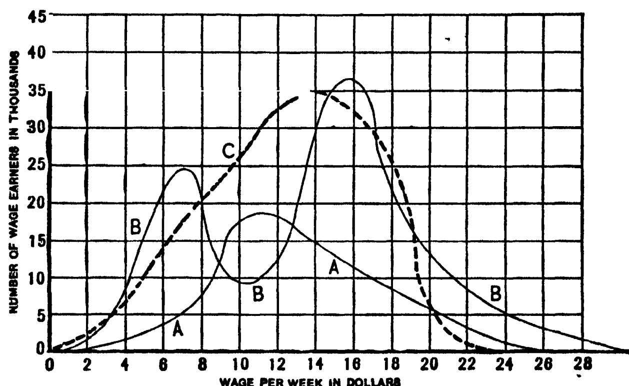


FIG. 8.

wages but the men tend to receive high wages. There are fewer workers receiving very high wages in state *C* than in either of the other two, but the general standing of wages is high.

It will be observed that, in Fig. 8, the comparison of

wages is partially obscured by the difference in the general altitude of the curves due to the fact that one state has so many more workers than another. To eliminate this defect, it is advisable to reduce all frequencies to percentages by dividing each class by the total number of items as has been done in the following table. Thus, in state *A*, the number of wage earners in

TABLE VII.

FREQUENCY TABLE SHOWING COMPARATIVE WAGES IN STATES *A*, *B*, AND *C*. BOTH ABSOLUTE AND PERCENTAGE FREQUENCIES BEING GIVEN.

Wages per Week.	State *A*.		State *B*.		State *C*.	
	No. of Wage Earners.	Percentage of Wage Earners.	No. of Wage Earners.	Percentage of Wage Earners.	No. of Wage Earners.	Percentage of Wage Earners.
$ 0– 1.99	25	0.0	210	0.1	1,114	0.5
2.00– 3.99	1,460	1.5	4,630	2.3	4,986	2.3
4.00– 5.99	3,784	3.9	16,424	8.1	10,102	4.8
6.00– 7.99	5,025	5.1	24,898	12.3	17,170	8.1
8.00– 9.99	13,200	13.5	12,122	6.0	22,054	10.4
10.00–11.99	17,420	17.7	8,964	4.4	28,402	13.3
12.00–13.99	16,142	16.5	17,220	8.5	33,960	15.9
14.00–15.99	13,240	13.5	35,116	17.3	34,817	16.3
16.00–17.99	10,940	11.2	34,963	17.2	31,460	14.8
18.00–19.99	7,964	8.1	17,842	8.8	24,972	11.7
20.00–21.99	4,982	5.1	12,240	6.1	3,417	1.6
22.00–23.99	2,786	2.8	7,963	3.9	546	.3
24.00–25.99	962	1.0	6,241	3.0		
26.00–27.99	70	.1	3,196	1.5		
28.00–29.99			971	.5		
Total	98,000	100.0	203,000	100.0	213,000	100.0

each class is divided by the total, 98,000, giving the percentages shown in the third column. The sum of the percentages must, of course, be 100 in each case. In Fig. 9, we see the results of plotting these **percentages**

PERCENTAGE HISTOGRAMS SHOWING COMPARATIVE WAGES IN STATES *A*, *B*, AND *C*.

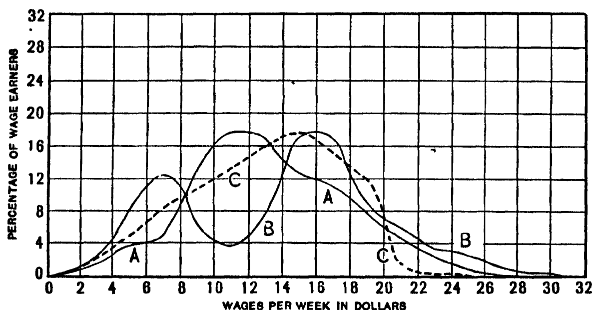


FIG. 9.

in the form of histograms. By this means, all states, large or small, are placed on an equal footing, and the study is much simplified. In most instances, these **percentage histograms** give much more satisfactory comparisons than do those histograms using the absolute figures.

In general, it may be said that smoothed histograms form one of the simplest and best of all means of comparing the frequency distribution of two or more groups of data and they are applicable to a great variety of cases.

Sec. 63. Cumulative Frequency Tables.

A cumulative frequency table is constructed by simply adding together the frequencies given in a simple frequency table. In this procedure, each class is made to include all the lower classes. If we return to our table of cornstalk heights and cumulate the frequencies, the results will be as follows:

TABLE VIII.

CUMULATIVE FREQUENCY TABLE SHOWING HEIGHTS OF CORN.

Height in Ft. (Size of Item). m	No. of Stalks (Frequency). f	Cumulative Frequency.
3– 4	3	3
4– 5	7	10
5– 6	22	32
6– 7	60	92
7– 8	85	177
8– 9	32	209
9–10	8	217
	$n = 217$	

Sec. 64. The Ogive.

By plotting the data given in the last column, we shall obtain a cumulative frequency graph or ogive. It will be remembered that in constructing a frequency polygon the frequency must be plotted at the **midpoint** of the class but, in laying out an ogive, it must always be plotted at the **upper limit** of the class instead. **Fig.** 10 shows such an ogive constructed for the above table and then smoothed. A glance at the figure shows how

much more readily this is accomplished in the case of the ogive than in that of the histogram. This is one of the marked advantages of the ogive.

ANGULAR AND SMOOTHED OGIVE SHOWING HEIGHTS OF CORNSTALKS.

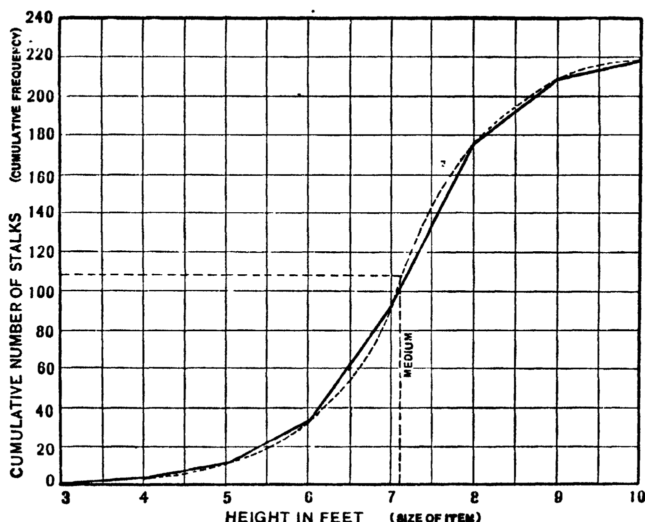


FIG. 10.

Ogives, like histograms, may be used for comparing groups of statistics in which time is not a factor. Just as in the case of histograms, better results are obtained if the frequencies are reduced to percentages. Ogives,

in general, are more difficult for the ordinary person to interpret than histograms and are used primarily for the determination of medians, quartiles, percentiles, etc., a matter which will be discussed in a later chapter.

Sec. 65. **General Rules for Construction of Graphs.**

1. Rule the axes in heavy black lines.

2. Choose a scale which will include all your items and at the same time fit the paper

3. In arranging the scale, always place round numbers on the heavy lines of your section paper. If the paper is divided on the decimal plan, number your scale 5, 10, 15 . . . ; 10, 20, 30 . . . , or in some other numbers which are readily divisible on the same scale as the paper, never using such a scale as 3, 6, 9 . . . , the fractions of which fail to correspond to the divisions on the paper. Never number the scale simply to agree with the frequencies given in the table.

4. Graphs should, in general, cover the main part of the sheet of paper used. They should be on a large enough scale to bring out such details as are desired, but a graph small enough to be taken in at a glance is preferable, for most purposes, to one of greater size.

5. The graphs should be as accurate as convenient to make them but, if they illustrate sufficiently the points that it is desirable to bring out, great precision in every detail is not essential.

REFERENCES.

ELDERTON, W. PALIN and ETHEL M. *Primer of Statistics.* Adam and Chas. Black, London, 1910. Chaps. I and II.

YULE, G. U. *Introduction to Statistics*, Chap. VI.

THORNDIKE E. L. *An Introduction to the Theory of Mental and Social Measurements.* Science Press, N. Y., 1904. Chaps. III and IV.

BOWLEY, A. L. *Elements of Statistics*, Chap. VII.

BOWLEY A. L. *The Measurement of Groups and Series.* Chas. and Edwin Layton, London, 1903. First Lecture.

EDGEWORTH, F. Y. *The Law of Error.* Ency. Britannica.

CHAPTER XII.

TYPES AND AVERAGES.

Sec. 66. **Uses of Types or Averages.**

Averages are used 1. **To give a concise picture of a large group.** We could not grasp the idea well if given the height of every tree in a forest but the average height is something definite and comprehensible.

2. **To compare different groups by means of these simple pictures.** It follows as a corollary of No. 1 that, before we can compare two groups, we must have a definite picture of each in mind; thus, two forests can only be compared by means of totals or averages of some sort.

3. **To obtain a picture of a complete group by the use of sample data only.** It has been found, in practice, entirely superfluous to measure the heights of every person of a race to obtain the typical height of that people. An average obtained from a few hundred samples is so close to the exact average of the whole that the difference is negligible.

4. **To give a mathematical concept to the relationship between different groups.** We may say that the trees in one forest are taller than in another but in order to find any definite ratio of heights it is necessary to resort to averages.

I. THE MODE.

Sec. 67. The Mode Defined.

One of the most useful of the types or averages is the mode. It is variously defined as **the most frequent size of item, the position of greatest density,** and **the position of the maximum ordinate in a smoothed histogram.** When we speak of the average man, the average income, etc., we usually mean the modal man or the modal income. We might say that the modal workingman's house contained five rooms, the modal contribution to a church collection is five cents, meaning, in each instance, that this is the vogue, the most usual occurrence, the common thing.

Sec. 68. Methods of Determining the Mode.

Once a smoothed histogram has been correctly constructed, the mode may usually be located, at a glance, by finding the size of item corresponding to the greatest ordinate or the highest part of the curve. It is perfectly possible for a histogram to have a number of distinct modes and, when this is true, each is located in the same way.

If there is but a single well-defined mode, the class containing it is at once located in the frequency table, but, in many instances, there are numerous irregularities in the table, though but one mode is really existent, and then the modal class is not so easily selected. In such cases, it is best to approximately locate the mode by **a process of grouping.** The procedure is as follows:

First, the frequencies are grouped by twos. Then, the upper limit of the group is shifted one space and again the frequencies are grouped by twos. Next, the grouping is done by threes. The upper limit is shifted down one space and the process repeated. The limit is then shifted another space and the process again repeated. If necessary, the grouping is now done by fours. This method is continued until regularity is secured and the point of maximum frequency is not changed by a shift in the upper limit of the group. The mode now lies in each of the largest groups in the later series of groupings used. In this way, it can be definitely placed within certain limits. The following table illustrates the process. Only the items near the mode are used, since the inclusion of the extremes is manifestly useless.

TABLE IX.

LOCATION OF MODE BY GROUPING.

Size of Item. *m*	Frequency. *f*					
5	48	100		156		
6	52		108		168	
7	56	116				178
8	60		122	182		
9	62	122			180	
10	60		118			174
11	58	114		177		
12	56		119		179	
13	63	123				171
14	60		108	148		
15	48	88			120	
16	40		72			
17	32					

In the first grouping, the mode apears to be either 13 or 14, since 123 is the maximum sum of frequencies obtained. In all the later groupings, however, the maximum sum is seen to be shifted to the neighborhood of 9. As the limits of the groups are changed, we find that 9 is the size of item whose frequency is **constantly contained** in the **maximum** group. This is true of no other size of item, hence the mode is located as approximately 9.

If the class interval used in a frequency table is large, it is often desirable to locate the mode **within** the limits of the class. We see that, in the rectangular histogram shown in Fig. 7, the class 6–7 is much larger than the class 8–9, hence the probabilities are that the exact location of the mode will be nearer 7 than 8. An empirical method which has proved serviceable is to locate the mode in the modal class according to the weights of the classes adjacent to it. This may be expressed by the following formula:

Let l = the lower limit of the class.
c = the class interval.
f_1 = the number of items in the next lower class.
f_2 = the number of items in next higher class.
Z = mode.

Then

$$Z = l + \frac{f_2 c}{f_2 + f_1}.$$

If we substitute in the formula the data from the

frequency table of cornstalk heights given in Sec. 61, we obtain the following:

$$Z = 7 + \frac{32 \times 1}{32 + 60}$$
$$= 7 + \frac{32}{92}$$
$$= 7.347+ \text{ the required mode.}$$

In exceptional cases in which the frequencies are very irregular, it may be advisable to use two or more classes on each side of the modal class as weights, but this should not be done if the graph is likely to be really multimodal.

The mode may also be located very roughly on an ogive by finding the point on the curve at which it is most nearly vertical. This may be accomplished mechanically be slipping a ruler along the ogive, keeping it constantly tangent to the curve and noting the moment when the angular movement of the ruler tends to reverse its direction. The ogive, however, is usually almost valueless in determining the mode.

Sec. 69. **Advantages of the Mode as a Type.**

1. **The mode is useful in cases in which it is desirable to eliminate extreme variations.** A few trees in the forest or a thousand-dollar check in the church collection would in no way disturb the mode but would affect the arithmetic average.

2. **In determining the mode, it is unnecessary to**

know anything about the extreme items except that they are few in number. We need have no record of the number of millionaires and the size of their estates or the number of paupers in order to find the modal wealth of the people of the United States.

3. **It may be determined with considerable accuracy from well-selected sample data.**

4. **It is the type that, to the ordinary mind, seems best to represent the group.** It is more intelligible to say that the modal wage of workingmen in a community is $2 per day than to say that the average wage is $2.17 when not a single man actually receives the latter amount.

Sec. 70. **Disadvantages of the Mode as a Type.**

1. **In many cases, no single, well-defined type actually exists.** One could scarcely picture a modal size of city which would mean anything. In wage statistics, we are likely to find several distinct modes corresponding to the various grades of labor.

2. **The mode is not at all useful if it is desirable to give any weight to extreme variations.** If one wished to learn how much wealth each person would have were all goods equally distributed, he would not be assisted by a knowledge of the present modal wealth.

3. **It cannot be located by any simple arithmetic process and, in many cases, is difficult to determine accurately by any method.**

4. **The product of the mode by the number of items**

does not give the correct total, as is the case when the arithmetic average is used.

5. **The mode may be determined by a comparatively small number of items of uniform size in a large group of varying size.** Thus, in a community having great variations in wealth, the modal value of possessions might be $992 simply because three people were listed at that amount while the wealth of all others varied. This difficulty is overcome in practice by using classes of considerable breadth.

II. THE MEDIAN.

Sec. 71. Defining and Locating the Median.

If a number of similar objects are placed side by side in order of their size, they are said to be **arrayed.** We have, in Fig. 5, lines representing an **array** of the lengths of 113 leaves. If any group of objects is thus arrayed, the middle one is known as the median item. Thus, in the leaf lengths illustrated, the fifty-seventh item would be the median, for it would have fifty-six items on each side of it. The median leaf-length, therefore, is about 6.7 cm. If there is an even number of items the median item does not actually exist, but it is assumed to be located between the two middle items. Were we to experiment with a much larger number of leaves, but chosen, as were the 113 in the illustration, purely at random, we should find that the length of the median leaf would, like that of the

modal leaf, remain practically constant. The median therefore proves a useful type to represent a given set of items.

The median may also be defined as that item whose size corresponds most closely to the size of all the other items in the array. Stated in mathematical language this means that, **when all deviations are considered positive,** the sum of the deviations from the median is a minimum.

DEVIATIONS FROM MEDIAN.

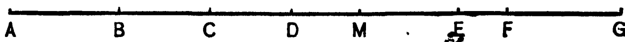


FIG. 11.

Demonstration (see Fig. 11):

Let AB, AC, AD, AM, AE, AF, and AG represent an array of seven items of which AM is evidently the median.

Then, the sum of the deviations from M is equal to $BM + CM + DM + ME + MF + MG$. Now let us take any other item as AE. The sum of the deviations from AE is equal to

$$(BM+ME)+(CM+ME)+(DM+ME)+ME+(MF-ME)+(MG-ME)=BM+CM+DM+MF+MG+2ME.$$

But this sum is greater by ME than the deviations from the median, hence the sum of the deviations (taken positively) from the median is a minimum.

When the items have been arrayed as in Fig. 5 the median is located simply by counting until the middle item is reached. In practice, however, we usually have the items given in a frequency table with a large num-

ber of items in each class. A simple way of finding the median in such cases is to first plot the data as an ogive and smooth. A horizontal line is now drawn through the midpoint of its altitude or its projection on the vertical axis. At the intersection of this horizontal line with the ogive, a vertical line is dropped to the horizontal axis and the point of intersection indicates on the scale the median required. In Fig. 10, there are 217 items, therefore, the horizontal line is drawn through the point representing 108.5 on the vertical axis and the median height is found to be 7.15 ft.

It is often desirable to definitely locate the median within a class by using the frequency table direct. This may be done by interpolation, the assumption being that the size of items varies uniformly throughout the class. If this assumption is true, the following formula will hold good.

Let M = the median.
c = the class interval of the class containing the median.
l = the lower limit of the class.
f = the number of items in the class.
i = the number of items up from the lower limit of the class at which the median item occurs.

Then

$$M = l + \frac{c(2i - 1)}{2f}.$$

10

$$M = l + \frac{N/2 - F}{f} \times c$$

In the frequency table given in Sec. 63, there are 217 items. The median item, then, is number 109 in the array. But the 109th item is the 17th item up in the array of 85 items in the 7–8 ft. class .Therefore $l = 7$, $c = 1$, $f = 85$, and $i = 17$. Substituting in the formula,

$$M = 7 + \frac{1(2 \times 17 - 1)}{2 \times 85}.$$

$M = 7.19$ ft., approximately the same result obtained by the graphic method.

Sec. 72. Advantages and Disadvantages of the Median as a Type.

The advantages may be enumerated as follows.

1. **It may usually be located with greater exactitude than the mode.** This is especially true in groups in which the mode is ill-defined.

2. **It is but slightly affected by items having extreme deviations from the normal.** In this respect, it resembles the mode more closely than it does the arithmetic average. The thousand-dollar check in the church collection does not affect the mode at all and it affects the median only as much as any other single item larger than the median would do, that is, the weight of this deviation is not increased by its extraordinary size, but the item receives the same weight as any other instance and no more.

3. **Its location can never depend upon a small number of items, as is sometimes the case with the mode.**

4. **If the number of the extreme items is known, their size is not required in determining the median.**

Thus, if the number of persons possessing over $100,000, and the number of paupers are known, the median of wealth could be calculated from statistics of the possessions of the intervening classes without considering the value of the property of either the extremely poor or the extremely rich.

5. **The median is especially useful for considering data, the items of which are not susceptible of measurement in definite units.** It is impossible to measure in specific units the mental characteristics of a child but it is perfectly possible to array a group of children according to their respective mentality. An arithmetic average, in such cases, is meaningless and practically useless for comparative purposes but a median can be legitimately determined and its characteristics compared with other similar medians.

Some of the disadvantages of the median are:

1. **Like the mode, it is not so readily determined by a simple mathematical process as is the arithmetic average.**

2. **As in the case of the mode, a correct total cannot be obtained by multiplying the median by the number of items.**

3. **Like the mode, it is not useful in those cases in which it is desirable to give large weight to extreme variations.**

4. **Unlike the mode, but like the arithmetic average, it is frequently located at a point in the array at which**

actual items are few. Thus, the median wage might accidentally fall on the $2.37½ per day while perhaps only a very few men actually received this amount.

5. **In a discrete series in which the items are so slightly dispersed that they fall largely in the modal class, there may be so many items of the same size as the median that it becomes very indefinite.** In such a case, the number of items larger than the median may be very different from the number of items smaller than the median. When this difference is too great, the value of the median as an average is largely destroyed.

On the whole, the median is one of the most valuable types for practical use, and for studies such as wages, distribution of wealth, etc., is often decidedly superior to either the mode or the arithmetic average.

III. THE ARITHMETIC AVERAGE OR MEAN.

A. The Simple Arithmetic Average.

Sec. 73. **Definition of the Arithmetic Average.**

The sum of all the items in a group is known as the **aggregate.** The arithmetic average may be defined as **the sum or aggregate of a series of items divided by their number.**

In computing the arithmetic average from a frequency table, the student must, of course, remember to multiply each item by its frequency before summating. When the size of items is only approximately

known, the midpoint of the class may usually be taken to represent each of the items therein without introducing any serious error. This is especially true when the class interval is small. One of the characteristics of the arithmetic average which is derived from the definition given above is that **the sum of the deviations (signs considered) of all the items therefrom equals zero.** This may be proved as follows:

Demonstration.

In Fig. 12, let KM_1, KM_2, and KM_3 be part of a series

DEVIATIONS FROM ARITHMETIC AVERAGE.



FIG. 12.

of items whose arithmetic average is KA. Then, by definition,

$$\frac{KM_1 + KM_2 + KM_3 \cdots KM_n}{n} = KA.$$

By substitution,

$$\frac{(KA - AM_1) + (KA - AM_2) + (KA + AM_3) \cdots + (KA + AM_n)}{n} = KA.$$

$$\therefore\ KA - AM_1 + KA - AM_2 + KA + AM_3 \cdots + KA + AM_n = n\overline{KA}$$

and

$$n\overline{KA} - AM_1 - AM_2 + AM_3 + \cdots AM_n = n\overline{KA}.$$

Therefore,

$$-AM_1 - AM_2 + AM_3 + \cdots AM_n = 0.$$

Hence, the sum of the deviations from the arithmetic average (signs considered) equals zero.

Sec. 74. Determination of the Arithmetic Average by the Short-cut Method.

If a group of large items are all of nearly the same size, it is frequently a time-saver to find the arithmetic average by the application of the following rule: **Assume any number as the average; find the sum of the deviations therefrom, having regard for the signs in each case; divide the sum of the deviations by the total number of items; then, add the quotient to the assumed average. The result is the true average.** To illustrate:

TABLE X.

SHORT-CUT METHOD OF COMPUTING THE ARITHMETIC AVERAGE.

Items.	Assumed Average.	Deviations from Assumed Average.
747	740	+ 7
742	740	+ 2
735	740	− 5
738	740	− 2
730	740	−10
736	740	− 4
		Total, −12

The number of items is 6.

$$-12 \div 6 = -2,$$

$$740 + (-2) = 738 = \text{the true average.}$$

In using the short-cut method in a frequency table,

it is, of course, necessary to multiply each deviation by the number of items in the group represented.

Algebraic Proof of the Short-cut Method.

In Fig. 13, let KM_1, KM_2, KM_3, and KM_4 be four instances of a series of n items and KA be the arithmetic average of the whole. Assume that the quantity

SHORT-CUT METHOD FOR ARITHMETIC AVERAGE.

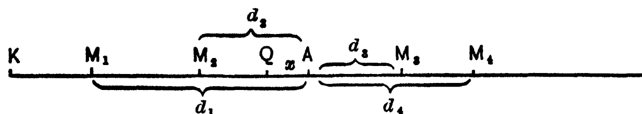


FIG. 13.

KQ is the arithmetic average. Let d_1, d_2, d_3, $\cdots$ d_n be the respective deviations of the items from the true arithmetic average KA. Let $QA = x$.

To prove:

$$\frac{\Sigma \text{ deviations from } KQ}{n} + KQ = KA.$$

Proof:

$$\frac{\Sigma \text{ deviations from } KQ}{n}$$

$$= \frac{-(d_1-x)-(d_2-x)+(d_3+x)+(d_4+x)\cdots+(d_n+x)}{n}$$

$$= \frac{-d_1-d_2+d_3+d_4+\cdots d_n+nx}{n} = \frac{nx}{n} = x,$$

since the sum of the deviations from the arithmetic

average = 0. But $x + KQ = KA$,

$$\therefore \frac{\Sigma \text{ deviations from } KQ}{n} + KQ = KA.$$

Q.E.D.

Sec. 75. **Advantages of the Arithmetic Average as a Type.**

1. Unlike the median or mode, it may be definitely located by a simple process of addition and division, and it is unnecessary to draw diagrams or arrange the data in any set form or series.

2. It gives weight to extreme deviations which is desirable in certain cases.

3. Unlike the mode, it is affected by every item in the group, and its location can never be due to a small class of items.

4. It is familiar to everyone and hence needs no explanation when used. The same cannot be said of the median or mode.

5. It may be determined when the aggregate and the number of items are known and information concerning the various items is entirely lacking. If we know the amount of sugar manufactured and imported into the United States annually and the population of the United States, we may calculate the average consumption of sugar per capita and never know how much any single consumer uses. This would be impossible with any other average,

Sec. 76. **Disadvantages of the Arithmetic Average as a Type.**

1. **It cannot be located on a frequency graph when such is already at hand.**

2. **It cannot be accurately determined where the extremes of a series are missing.** In this respect, it is surpassed as a type by both the median and the mode.

3. **It emphasizes the extreme variations which in most cases is undesirable.**

4. **It cannot like the median be used with advantage in the study of incommensurable quantities.**

5. **It is likely to fall where no data actually exist.** It is easy to find by computation that the average number of persons in a family is 5.41, although such a number is evidently impossible.

B. The Weighted Arithmetic Average.

Sec. 77. **Definition of the Weighted Average.**

By a weighted average, we mean one whose constituent items have been multiplied by certain weights before being added, the sum thus obtained being divided by the sum of the weights instead of by the number of items. The weights used may represent the actual or estimated number of items existing in a certain group, in which case it does not differ essentially from a simple average. If, for example, we know the wages paid to a few men in each occupation in an industry and we desire to ascertain the average wage for that

industry, we must multiply the average wage found for each occupation by the number of men engaged in that occupation, summate the results, and divide the sum by the total number of men employed in the industry. If we took simply an arithmetic average of the samples, it would evidently be inaccurate unless the numbers of samples in different occupations were in the same ratio as the total number of men engaged in the respective occupations. The result obtained by means of the weighted average is, in this case, approximately the same as if we had the wage of each man in the industry recorded and then obtained a simple arithmetic average of the entire data.

In other instances, the weights do not represent numbers but stand for estimates of relative importance. In making up a semester average of the grades received by a student, the teacher usually assigns an arbitrary weight to the different factors as, for example, 3 to the class grade, 2 to written work, and 4 to the final examination, the sum of the products being divided by 9. In this case, the weighted average corresponds less closely to any simple average.

Sec. 78. **Effects of Weighting.**

If the number of weights used is small, the size of weights chosen is likely to have a marked effect on the average. When the weights are very numerous, the chances are that they will tend to offset each other, so that the results will be but little different from those

obtained by using a simple average. This depends largely, however, on whether there is any relationship between the size of the weight and the size of the item. If large weights go with large items, or vice versa, the average will be seriously affected whenever weights are neglected or erroneous ones used. One usually finds in the study of wages that low wage men are numerous and high wage men few. If, then, one used a simple average of the wages in all the occupations of an industry, the single superintendent would be given as much weight as the thousand common laborers under his direction and his large salary would make the average wage appear much too large. In such cases, therefore, weights cannot be neglected. It may, however, be mathematically demonstrated[1] that an error in weights tends to be much less serious in its effects on the final result than an error in the size of the original items.

An error in the size of the original items cannot be remedied by adjustments in the weights used. Hence, the following general rule may be enunciated:

The items should be as exact as possible and the weights used should be approximately accurate but great exactness in the size of weights causes much extra work and is unnecessary.

[1] For proof see Bowley's *Elements of Statistics*, pp. 203–205.

IV. THE GEOMETRIC AVERAGE.

Sec. 79. **Definition of the Geometric Average.**

The geometric average is obtained by multiplying together the n items in a series and then extracting the nth root of the product. It is almost necessarily computed by the use of logarithms. It was largely used by Jevons in his study of prices, but has not found much favor among statisticians in general.

Sec. 80. **Characteristics of the Geometric Average.**

The geometric average is always slightly smaller than the arithmetic average. **It gives comparatively little weight to extreme variations.** In this respect, it lies between the arithmetic average and the median. **It requires more time to compute than other averages. It has the disadvantage of not being commonly understood and being somewhat difficult of comprehension to the non-mathematical mind.**

REFERENCES.

Elderton, W. P. and E. M. *Primer of Statistics*, Chaps. I and II.
Bowley, A. L. *Elements of Statistics*, Chaps. V and VI.
Yule, G. U. *Introduction to Statistics*, Chap. VII.
Bailey, Wm. B. *Modern Social Conditions*, pp. 31–40.
Bowley, A. L. *Elementary Manual of Statistics*, Chap. III.
Bertillon, Jacques. *Cours Élémentaire de Statistique*, Chap. IX.

CHAPTER XIII.

DISPERSION.

Sec. 81. Explanation of Dispersion.

The term **dispersion** is used to indicate the fact that, within a given group, the items differ from one another in size, or, in other words, that there is a lack of uniformity in their magnitudes. When we say that the dispersion is slight, we mean that this difference is trivial when compared with the absolute size of the average item while the dispersion is said to be great when such variation is relatively large. If, for example, a military company were composed entirely of men ranging in height from 68 to 70 inches, we would say that their height was very uniform or that the dispersion was slight. If, however, the shortest men were only 62 inches and the tallest were 74 inches in stature we should then say that there was considerable dispersion in the heighths of the men. To cite another instance: In the early days of the frontier, wealth was quite evenly distributed but, today, with our millionaires and paupers, we have a wide dispersionof wealth. In Array I of Fig. 14, we find a total dispersion in length of five-sixteenths of an inch, while in Array II the dispersion is ten-sixteenths of an inch.

The dispersion of a group may be measured by the difference in size or characteristics of the most extreme

ARRAYS OF LEAF LENGTHS ILLUSTRATING DISPERSION.

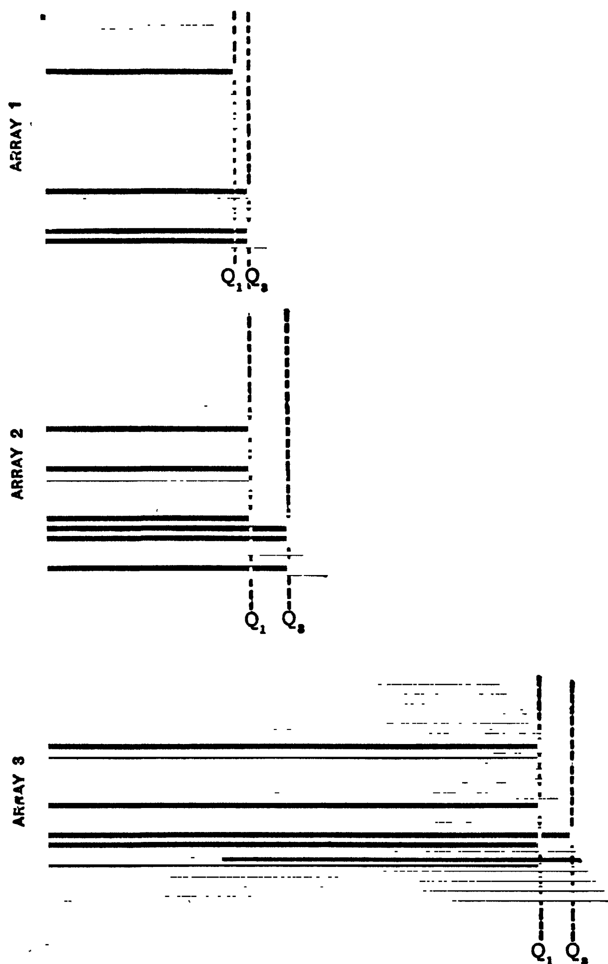


FIG. 14.

items, in other words, the **range**, or it may be measured by the general deviation of the items from the type. The **range** is too indefinite to be used as a practical measure of dispersion. If, in a community, the shortest adult were 5 ft. and the tallest 6 ft. 1 in. in height, the range would evidently be 13 inches but, if a dwarf whose height was but 3 ft. 6 in. should move into the neighborhood, the range would suddenly be increased to 31 inches, while the average height of the people would be but trivially affected. It is evident that a measure so radically affected by stray items at the extremes must be practically valueless. We must, therefore, measure dispersion by the deviation from some type or average or at least modify the range in such a way as to eliminate the scattering extreme items.

Dispersion may also be measured absolutely or relatively. In the first case, the average size of the items makes no difference, while, in the second case, this is of fundamental importance. A difference of an inch in the heights of a company of men would be very slight but a difference of an inch in the lengths of their noses would be decidedly noticeable. In Fig. 14, the absolute range of dispersion in Arrays II and III is the same in each case, but, relatively, it is more than twice as great in Array II as in Array III for the average size of an item in Array III is more than double that of one in Array II,

To render the **relative** dispersion in different groups comparable, it is necessary to obtain a coefficient of dispersion for each group. **This coefficient is computed by dividing the absolute measure of dispersion used by some quantity representing the typical sized item.** The coefficient, then, represents the **fraction** of variation occurring generally in the given group of data.

Sec. 82. **Moments.**

Dispersion is commonly measured by finding the average deviation of the items from some one of the types, usually the arithmetic average, the mode or the median. For measuring such deviations, the various **moments** are used. The **first moment** is simply the average deviation, or, in other words, the sum of the deviations divided by the number of items. If m_1, m_2, m_3, $\cdots m_n$ are the items, n the number of items, and d_1, d_2, d_3, $\cdots d_n$ the respective deviations of the items from the type, then the moments are expressed as follows:

First moment: $\Sigma d/n$.

Second moment: $\Sigma d^2/n$.

Third moment: $\Sigma d^3/n$.

I. MEASURES AND COEFFICIENTS OF DISPERSION.

A. FIRST GROUP. BASED ON FIRST MOMENT.

Sec. 83. **The Average Deviation and the Corresponding Coefficient of Dispersion.**

In computing the average deviation, all deviations

are considered positive. They may be computed either from the mode, the median or the arithmetic average.

If $m_1, m_2, m_3, \cdots m_n$ are the items, n in number, having respective deviations of $d_1, d_2, d_3, \cdots d_n$ from the given type, the average deviation being represented by δ, and if a is the arithmetic average, M the median, and Z the mode, the average deviation may be computed by either of the following formulæ according to the average used.

The average deviation from the arithmetic average,

$$\delta = \frac{\Sigma(m - a)}{n} \text{ or } \frac{\Sigma d}{n}.$$

The average deviation from the median,

$$\delta_M = \frac{\Sigma(m - M)}{n} \text{ or } \frac{\Sigma d_M}{n}.$$

The average deviation from the mode,

$$\delta_z = \frac{\Sigma(m - Z)}{n} \text{ or } \frac{\Sigma d_z}{n}.$$

These measures of dispersion may be reduced to coefficients by dividing each by the respective average employed. The coefficient of dispersion based on the arithmetic average

$$= \frac{\Sigma(m - a)}{na} \text{ or } \frac{\Sigma d}{na} \text{ or } \frac{\delta}{a}.$$

That based on the median

$$= \frac{\Sigma(m - M)}{nM} \text{ or } \frac{\Sigma d_M}{nM} \text{ or } \frac{\delta_M}{M}.$$

That based on the mode

$$= \frac{\Sigma(m - Z)}{nZ} \text{ or } \frac{\Sigma d_z}{nZ} \text{ or } \frac{\delta_z}{Z}.$$

The following example illustrates the mode of computing from a frequency table this coefficient of dispersion, the deviations from the median being used as a basis.

TABLE XI.

COMPUTATION OF THE AVERAGE DEVIATION.

Size of Item. m	Frequency. f	Deviation from Median. d_M	fd_M
4	2	3	6
5	3	2	6
6	5	1	5
7	8	0	0
8	6	1	6
9	4	2	8
10	2	3	6
11	1	4	4
	$n = 31$		$\Sigma d_M = 41$

The median $= M = 7$.

$$\delta_M = \frac{41}{31} = 1.32 +$$

The coefficient of dispersion

$$= \frac{\delta_M}{M} = \frac{1.32}{7} = 0.19 -$$

In the above example, the median has been considered a whole number which is correct only if the series is discrete. If the series is continuous, it is necessary to interpolate in the fourth class to locate it exactly. It must also be noted that all deviations are treated as positive.

The characteristics of this coefficient of dispersion are:

1. It is easy to compute and comprehend.
2. It takes every item into consideration.
3. It gives weight to deviations according to their size, extreme deviations having more weight than small ones, but not being disproportionately magnified.

This coefficient is a good one to use in many economic studies as, for example, in calculating the personal distribution of wealth in a community or a nation, since the very rich and the very poor are both taken into account. The question as to which average should be used in the computation is not usually of great importance. For the distribution of wealth, it is probably preferable to use the deviations from the median.

B. Second Group. Based on the Second Moment.

Sec. 84. **The Standard Deviation and Coefficient.**

The only measure of dispersion in this group in extensive use at present is the standard deviation. It is conceivable that a similar measure might be used whose deviations were based upon the mode or median, but

the standard deviation is invariably computed from the arithmetic average. The formula is as follows:

σ = standard deviation.

The other letters are used as in Sec. 83. Then

$$\sigma = \sqrt{\text{Second Moment}} = \sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{\Sigma(m-a)^2}{n}}.$$

In calculating the standard deviation by use of an ordinary frequency table, the following illustrates the direct method.

TABLE XII.

CALCULATION OF STANDARD DEVIATION FROM A FREQUENCY TABLE: DIRECT METHOD.

Size of Items in Mms. m	Frequency. f	mf	Deviation. d	d^2	fd^2
8	2	16	−3	9	18
9	4	36	−2	4	16
10	6	60	−1	1	6
11	9	99	0	0	0
12	6	72	+1	1	6
13	4	52	+2	4	16
14	2	28	+3	9	18
	$n=33$	$\Sigma m=363$ $a=11$			$\Sigma d^2=80$

$$\sqrt{\frac{\Sigma d^2}{n}} = \sqrt{\frac{80}{33}} = \sqrt{2.4242} = 1.56 -.$$

Having obtained the standard deviation, all that is necessary to derive the corresponding coefficient of dispersion is to divide by the arithmetic average.

Therefore, the coefficient of dispersion

$$= \frac{\sigma}{a} = \frac{1.56}{11} = 0.14 +.$$

Sec. 85. **The Short-cut Method for Computing the Standard Deviation.**

The preceding method of computing the standard deviation is the simplest if the arithmetic average chances to be an even number. When, however, it is fractional, the effort involved in squaring and multiplying the decimals is considerable and it is preferable to use the short-cut method instead. The rule therefor is as follows: **Select some whole number approximating the arithmetic average; compute the deviations therefrom; square each; summate; subtract therefrom n times the square of the difference between this number and the true average; divide by n; extract the square root of the quotient.**

The algebraic formula employed in this method is:

If x = the assumed average

and a = the true average,

Then

$$\sigma = \sqrt{\frac{\Sigma(m - x)^2 - n(a - x)^2}{n}}.$$

To illustrate the method by example.

Assumed average = x = 9.

TABLE XIII.

COMPUTATION OF THE STANDARD DEVIATION BY THE SHORT-CUT METHOD.

Size of Item. m	Frequency. f	mf	$m-x$ or d_x	$(m-x)^2$ or d_x^2	fd_x^2
6	2	12	−3	9	18
7	4	28	−2	4	16
8	5	40	−1	1	5
9	7	63	0	0	0
10	4	40	+1	1	4
11	3	33	+2	4	12
12	1	12	+3	9	9
	$n=26$	$\Sigma m=228$			$\Sigma d_x^2=64$

$$a = \frac{228}{26} = 8.77-,$$

$$a - x = 0.23+,$$

$$(a-x)^2 = 0.053,$$

$$n(a-x)^2 = 1.375+.$$

$$\sigma = \sqrt{\frac{\Sigma d_x^2 - n(a-x)^2}{n}}$$

$$= \sqrt{\frac{64 - 1.375}{26}} = \sqrt{2.4086} = 1.55$$

The standard coefficient of dispersion then equals σ/a. But

$$\frac{\sigma}{a} = \frac{1.55}{8.77} = 0.177.$$

The correctness of the short cut method is based upon the following proposition: *The sum of the squares of the deviations from the arithmetic average is a minimum.* This theorem is demonstrated thus:

Given:

$$a = \text{the true arithmetic average.}$$
$$x = \text{any other assumed number}$$

To prove:

$$\Sigma(m - x)^2 > \Sigma(m - a)^2.$$

Proof:

$$(m - x)^2 = m^2 - 2xm + x^2,$$
$$\Sigma(m - x)^2 = \Sigma m^2 - 2x\Sigma m + nx^2.$$

But

$$\Sigma m = \text{the aggregate} = na.$$
$$\therefore \Sigma(m - x)^2 = \Sigma m^2 - 2xna + nx^2$$
$$= \Sigma m^2 + n(x^2 - 2ax)$$
$$= \Sigma m^2 + n(x^2 - 2ax + a^2) - na^2$$
$$= \Sigma m^2 - na^2 + n(x - a)^2.$$

But

$$\Sigma(m - a)^2 = \Sigma m^2 - 2a\Sigma m + na^2$$
$$= \Sigma m^2 - 2a \cdot na + na^2$$
$$= \Sigma m^2 - na^2.$$
$$\therefore \Sigma(m - x)^2 > \Sigma(m - a)^2. \qquad \text{Q.E.D.}$$

From the above, it follows that

$$\Sigma(m - x)^2 - n(x - a)^2 = \Sigma(m - a)^2.$$

$$\therefore \sqrt{\frac{\Sigma(m - x)^2 - n\,(x - a)^2}{n}} = \sqrt{\frac{\Sigma(m - a)^2}{n}}.$$

But

$$\sigma = \sqrt{\frac{\Sigma(m - a)^2}{n}}.$$

$$\therefore \sigma = \sqrt{\frac{\Sigma(m - x)^2 - n(x - a)^2}{n}}.$$

But this is the formula for the short-cut method for the standard deviation and it is thus proved correct.

Sec. 86. Characteristics and Uses of the Standard Deviation and Coefficient.

The standard deviation has, in the past, been used

more by biologists than by economists. The squaring of the large deviations gives more weight to extreme instances than to those differing but slightly from the mean and, for some purposes, this property is valuable. In most economic studies, however, the reverse tends to hold true and the average deviation is therefore preferable. An important exception to this rule, however, is found in the use of this coefficient in the computation of Karl Pearson's coefficient of correlation, a subject which will be discussed in a later chapter. The squaring of the deviations eliminates the negative signs and hence facilitates the mathematical manipulation of the figures. This is a valuable property of the standard deviation when it is used for advanced work in the study of symmetrical frequency distributions and, largely for this reason, it has proved a favorite with biologists. On the other hand, it requires considerably more effort to compute the standard deviation than the average deviation and, partially for this reason, the latter is commonly used by economists unless there is some special reason for preferring the former.

Another measure of dispersion based, like the standard deviation, on the second moment is the modulus, commonly represented by *c*. The formula for it is:

$$c = \sqrt{\frac{2\Sigma(m - a)^2}{n}} \text{ or } \sqrt{\frac{2\Sigma d^2}{n}}.$$

It has little place in the field of elementary statistics and so will not be discussed in this book.

C. THIRD GROUP—BASED ON QUARTILES.

Sec. 87. Quartiles, Deciles, etc.

The median was defined as the middle item of the array. Similarly the quartiles are those items that divide the number of items in an array into fourths, the deciles those that divide it into tenths, the percentiles into hundredths, etc. The second quartile, the fifth decile and the median are evidently synonymous. The median is the $\frac{n+1}{2}$ item; the first quartile is the $\frac{n+1}{4}$ item; the third quartile is the $\frac{3(n+1)}{4}$ item; the first decile is the $\frac{n+1}{10}$ item; the seventh decile is the $\frac{7(n+1)}{10}$ item; the twenty-fourth percentile is the $\frac{24(n+1)}{100}$ item; etc. In Fig. 5 we found the median to be the fifty-seventh item in the group. Similarly, the first quartile would be the twenty-ninth item, and the third quartile would be the eighty-fifth item. The quartiles, deciles, etc., are usually located by means of an ogive, its altitude being divided into fourths, tenths, etc., as the case may require. They may also be located in an array by simple division or in a frequency table by division and interpolation within a group, following the same formula used for determining the median.

Sec. 88. **The Quartile Measure and Coefficient of Dispersion.**

All the measures of dispersion previously discussed have taken into account the deviation of each particular item. The measure which we are now about to consider gives us a general idea of the dispersion of an array without going into so much detail. Half of the items in each array are included between the first and third quartiles. If the dispersion of this half of the items is fairly representative of the whole, we have, here, a very simple method of measuring it. In Fig. 5 the first quartile is about 6.3 cm. while the third approximates 7.1 cm., giving a range of fluctuation of 0.9 cm. The dispersion, however, if measured from the midpoint between the quartiles would be but half that amount or about 0.45 cm. In Arrays I and II in Fig. 14, we see the effect of a change in the amount of dispersion in a group on the distance between the quartiles. When the distance between the extreme variates is doubled the distance between the quartiles is approximately doubled also. The **quartile deviation** which is probably the simplest way of approximating the dispersion of an array has the following formula.

If

$$Q_1 = \text{the first quartile,}$$

and

$$Q_3 = \text{the third quartile,}$$

Then

$$\text{the quartile deviation} = \frac{Q_3 - Q_1}{2}.$$

The quartile deviations in Arrays II and III in Fig. 14 are equal but, since the average size of item in Array III is more than twice as large as in Array II, it is necessary to divide each by a quantity representing its typical size. This requirement would seem to be fairly fulfilled by the average of the quartile lengths or

$$\frac{Q_3 + Q_1}{2}.$$

The quartile coefficient of dispersion, then, would be

$$\frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}.$$

The quartile deviation and its coefficient have to their credit the merit of simplicity and ease of computation and are highly satisfactory if one is dealing only with the main body of an array and cares nothing about extreme variations. A glance at Fig. 5 will show that the length of all leaves shorter than the first quartile or longer than the third would have no effect whatever on the quartile deviation or coefficient. Yet, half the leaves fell outside these limits and some of the variations might have been very marked indeed. The quartile deviation is, therefore, useless when it is desired to give weight to the extremes in which respect it is exactly the opposite of the standard deviation; the average deviation occupying the intermediate and, for general purposes, superior position.

Sec. 89. **The Lorenz Curve.**

We have seen that graphic methods are very useful in illustrating frequency distribution by means of both absolute and percentage histograms. Another curve, worked out by Dr. Lorenz, illustrates very nicely the

Lorenz Graph, Showing Distribution of Wealth.

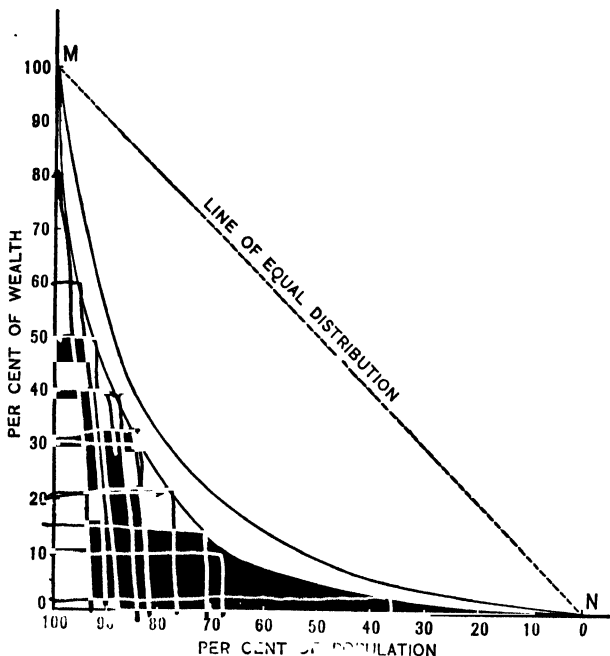


Fig. 15.

dispersion of a group and, in the study of the distribution of wealth, has proved especially applicable.

it does not, like a coefficient of dispersion, furnish a numerical measurement of the distribution of wealth and, in this respect, is inferior, but its particular merit lies in the fact that it pictures the distribution among the various sections of the population and does not merely give an average for the whole group. Fig. 15 outlines the mode of constructing this graph. If wealth were equally divided among the people, we should evidently have a straight line, like MN, connecting the extremes of the scales. In practice, we get curves like a or b. The closer the curve approaches the line of equal distribution the greater is the homogeneity of wealth indicated; the further it bows away from that line, the larger the percentage of the population in poverty and the greater the concentration in the hands of a few multimillionaires.

In the comparative study of different times or periods, we usually find that the curves tend to coincide near the extremities. When this occurs, it is well to plot a little of the extreme parts of the curves on separate sheets. In plotting the upper extremity, the horizontal scale may be greatly magnified, and when studying the right-hand extremity, the vertical scale may be correspondingly increased. In this way, the different curves are separated so that the variations at the extremes may be successfully analyzed.

This form of graph is also applicable to studies of the distribution among the population of land, wages,

income, etc. On the whole, it is more serviceable for these purposes than the percentage histogram and forms a valuable supplement to the coefficient of dispersion.

REFERENCES.

YULE, G. U. *Introduction to Statistics*, Chap. VIII.
ELDERTON, W. P. and E. M. *Primer of Statistics*, Chap. IV.
BOWLEY, A. L. *Measurement of Groups and Series*, Lects. II to IV.

CHAPTER XIV.

SKEWNESS.

Sec. 90. **Explanation of Skewness.**

By the term **skewness** as applied to frequency distributions we denote the **opposite of symmetry** indicating that the dispersion of the items within a given group is not symmetrical or, in other words, that, at points of equal deviation above and below the mode, the frequencies are unequal. Suppose, for illustration, that the wheat yields of all the farms of a certain county are tabulated and a frequency table constructed from the data thus collected. If the soil of the whole county is comparatively uniform and the modal yield of wheat is fifteen bushels per acre, we are likely to find the class-frequency in any pair of classes, on opposite sides of the mode and equidistant therefrom, to be nearly identical for, as we depart from the mode, the numbers of farms in the classes in which the yield is more than fifteen bushels will probably fall off in approximately the same ratio as the numbers in the classes producing less than fifteen bushels, thus giving us a normally symmetrical frequency distribution like that already studied in respect to the dice throws described in Sec. 60. If, on the other hand, there exists within the county in question a limited area of extremely sterile soil which is, nevertheless, utilized for wheat culture we should

find the production of this class of farms removed from the modal crop by an interval much greater than that separating the class containing the most fertile farms from the mode. In this case, the dispersion would no longer be symmetrical and the distribution would be said to be skewed toward the lower side.

The meaning of skewness is most easily made intelligible by the construction of a histogram. If skewness is present, the graph no longer presents the normal, symmetrical, bell-shaped form but the base is drawn out to a greater extent on one side than on the other as illustrated by graph *B* in Fig. 16, the lower part of

HISTOGRAMS ILLUSTRATING SKEWNESS.

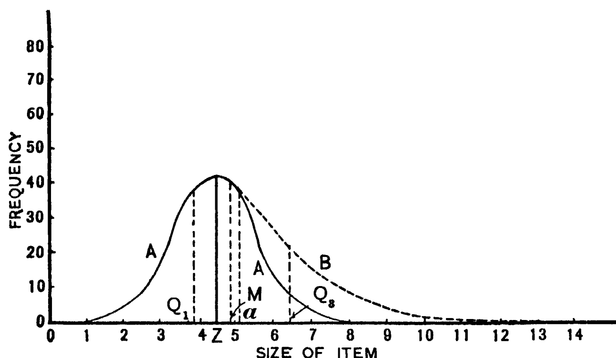


FIG. 16.

the histogram being here skewed far to the right from its normal position at *A*.

In the analysis of social phenomena, a perfectly symmetrical histogram is the exception and a large degree of skewness is to be frequently expected. Practical applications of the measurements of skewness have thus far, however, been confined largely to the biologic field but, in certain cases, such measurements are also useful in the study of economic statistics.

Sec. 91. **The Effect of Skewness on the Sequence of Averages.**

In the symmetrical histogram *A*, shown in Fig. 16, the arithmetic average, mode and median are all coincident at *Z*. In the curve, *B*, they have separated. The large items, far out to the right, while comparatively few in number, have had considerable effect in pulling the arithmetic average over in that direction for **it is always located at the center of gravity of the histogram** and, like weights hung far out on the long arm of a lever, these extreme instances prove more powerful than their mere numbers would indicate.

The median, **which bisects the area of the histograms,** is likewise shifted to the right by the accession of new instances on that side, but the size of these instances in this case gives them no added weight and the median, therefore, moves a lesser distance than does the arithmetic average. In curves not diverging too widely from the symmetrical form, the median usually travels over two thirds of the space covered by the arithmetic

average. Therefore, approximately,

$$M = Z + \tfrac{2}{3}(a - Z).$$

The mode, being in no wise affected by the addition of the new items, remains at its original location. We have, therefore, in a skew curve, a normal sequence of mode, median, and arithmetic average, the last being carried furthest in the direction in which the curve is skewed.

Sec. 92. **Measures and Coefficients of Skewness.**

If we desire to compare the skewness of one curve with that of another, it is necessary to reduce it, in every instance, to some numerical quantity. Measures of skewness must be reduced to coefficients for the same reason that measures of dispersion were so reduced but, in the case of skewness, the average size of item does not constitute a suitable divisor, for the question now is not how much the curve is skewed in proportion to the size of items involved, but how much more the items deviate on one side of the average than on the other. Hence, the denominator chosen must, invariably, be some measurement of the average deviation or dispersion of the items. With these points in mind, we shall proceed to consider some of the most commonly used measures and coefficients of skewness.

Sec. 93. **First Measure and Coefficient of Skewness.**

The distance that the arithmetic average is pulled beyond the mode makes one of the simplest possible measures of skewness.

If a = the arithmetic average,
M = the median,
δ_M = the average deviation from the median,
δ = the average deviation from the arith. average,
Z = the mode,
δ_Z = the average deviation from the mode,
j = the coefficient of skewness,
then the simplest measure of skewness is represented by the formula $a - Z$ and the simplest coefficient by

$$\frac{a - Z}{\delta_Z}, \quad \therefore j = \frac{a - Z}{\delta_Z}.$$

It makes little difference whether δ_Z or δ is used in the denominator, provided that the same one is employed in each case.

While the above is the ideal measure of skewness, the mode is often so ill-defined as to make it necessary to use as the numerator, instead, the difference between the median and the arithmetic average. As was noted in Sec. 91, this quantity is usually but one third as large as the difference between the mode and mean and this fact puts it at a disadvantage when the skewness is slight. When the median is employed, the formula for the coefficient becomes

$$j = \frac{a - M}{\delta_M}.$$

Sec. 94. **Second Measure and Coefficient of Skewness.**

This measure is based on the fact that, in a skew curve, the median no longer lies half way between the

quartiles, for the quartile nearest to the extended base of the curve is pulled in that direction more than the quartile opposite. This occurs because the quartile on the skew side is moving toward the region of *lesser* density or, in other words, lower frequency, thus having its movement accelerated, while the quartile on the opposite side is approaching the zone of *maximum* density and, hence, has its movement retarded. The median, lying usually near the mode, where the frequency is high, moves but slowly. The result is a gradual divergence in the relative distances of the quartiles from the median, and the difference in these two distances is utilized as a measure of skewness. The formula for this second measure of skewness, then, is $(Q_3 - M) - (M - Q_1)$ or $Q_3 + Q_1 - 2M$.

This is reduced to a coefficient by dividing by the quartile deviation. The coefficient thus obtained is:

$$j = \frac{Q_3 + Q_1 - 2M}{Q_3 - Q_1}.$$

This coefficient has the same weakness common to the quartile coefficient of dispersion—it fails to take into account the size of extreme variations. In calculating by this method the dispersion or skewness for a curve showing distribution of wealth in the United States, the result would be in no wise affected whether the ten thousand richest persons in the United States owned $100,000 or $100,000,000 each. In either case, they would all be far above the upper quartile and so the

location of both quartiles and median would be entirely unaffected. The merit of this coefficient lies in the fact that it is simple and easy to compute and is sufficiently accurate for practical purposes in those studies in which the extreme instances are not considered of fundamental importance.

Sec. 95. **Third Measure and Coefficient of Skewness.** This coefficient is based upon the third moment and depends upon the fact that cubing a quantity does not change its sign. It was demonstrated in Sec. 73 that the sum of the deviations (signs considered) from the arithmetic average equalled zero. In a skew curve, however, the sum of the **cubes** of the deviations from the arithmetic average does **not** equal zero for the process of cubing increases the relative importance of the extreme items and these are most important on the skewed side of the curve. The result is that the third moment itself furnishes a satisfactory basis for measure of skewness. It is, in one respect, the exact opposite of the quartile measure of skewness, since it emphasizes the extremes by which the other measure was not at all affected. The formula for this measure of skewness, then, is

$$\sqrt[3]{\frac{\Sigma(m-a)^3}{n}} \text{ or } \sqrt[3]{\frac{\Sigma d^3}{n}}.$$

To reduce this measure to a coefficient, various denominators may be used. Since it emphasizes the

extremes, it is well to use a denominator of the same type, such as the standard deviation in which case the formula would read.

Let j = the coefficient of skewness. Then

$$j = \frac{\sqrt[3]{\frac{\Sigma d^3}{n}}}{\sigma}.$$

The average deviation might be substituted instead, making the formula

$$j = \frac{\sqrt[3]{\frac{\Sigma d^3}{n}}}{\delta}.$$

This coefficient of skewness seems to be one of the best, but requires considerable work for its computation.

Various other measures and coefficients have been worked out but most of them are too complicated to be of practical value in elementary work.

REFERENCES.

YULE, G. U. *Introduction to Statistics*, Chap. VI.
BOWLEY, A. L. *Measurement of Groups and Series*, Lecture II.

CHAPTER XV.

HISTORICAL STATISTICS.

Sec. 96. **General Characteristics.**

In the preceding chapters, we have been dealing largely with data in which time is not a factor. One of the most important fields of statistics, however, is that which compares phenomena at different dates. This may be done in several different ways, among which the following are common.

1. Tables of absolute figures,
2. Absolute historigrams,
3. Logarithmic tables,
4. Logarithmic historigrams,
5. Index numbers,
6. Index historigrams.

We shall discuss these in order, omitting the first one, which scarcely needs explanation.

Sec. 97. **Absolute or Ordinary Historigrams — Smoothing — the Moving Average — the Trend.**

The numerical record of the changes of a variable during a number of successive intervals of time may be denominated **a historical series** and the graphs obtained when this historical series is plotted, using the sizes of the variable as ordinates and time intervals as abscissæ, is called a **historigram.** This must not be con-

fused with the histogram into the construction of which **time** does not enter.

The accuracy of a historical series, and, therefore, of a historigram depends manifestly on the length of time intervening between the records. An hourly record of temperature is more accurate than a daily, a daily than a weekly, etc.

In constructing a historigram, the original points plotted from the data may be connected by straight lines, but it is usually preferable to smooth the graph into a curve, the rules for smoothing being somewhat similar to those applying to histograms. In smoothing free hand, one should remember that the maximum possible radius of curvature should be constantly maintained, thus avoiding, as far as may be, all sharp breaks, unless such breaks are known to have actually occurred. In such variables as records of population, temperature, etc., changes rarely occur suddenly and sharp angles are therefore normally absent.

One of the best methods of smoothing certain varieties of historigrams is to use a **moving average** to obtain **a trend. It is only useful in those historigrams which manifest more or less periodicity and the object of using the moving average is to rid the historigram of these fluctuations. In determining on the size of groups to be used in calculating a moving average, one should use a period of time approximately equal to the length of the cycle which it is desired to eliminate.**

The best method of determining upon the proper length of cycle to use for the moving average group is to first plot the data as a historigram and then observe the average time-distance between the consecutive crests and between the successive troughs of the waves, this giving the approximate wave-length. The data given in the table below are plotted in Fig. 17. From this historigram, we see that the wave-length runs from six to eight days. It is preferable to use an odd number of days for the moving-average group, so that the average may be plotted opposite the central item of the group. In this case, then, we shall choose seven days as the appropriate length of period.

The first step in the computation is to obtain the average of the first seven items and place it opposite the fourth item. The average of the second to eighth items, inclusive, is next found and placed opposite the fifth item. This process is continued to the end of the series with the results shown in the table below. A shorter method, when the groups are large, is to add each time to the last total the difference between the number added and the number dropped.

In the table below, for example, the average from March 1 to 7 inclusive is 24.0°, from March 2 to 8, inclusive, is the same, since the same number 20° is both added and subtracted. For March 3 to 9, 28° is added and only 25° subtracted, hence the total of the group is increased by 3° and the average becomes

24.4°. This process is considerably facilitated by cutting a slot in a piece of cardboard just long enough to cover in the table the size of group desired.

It is impossible to accurately carry out a trend to the extremes of the data. The curve may be carried out free hand to each end or one may form artificial final groups by duplicating the number found at the

TABLE XIV.

TABLE ILLUSTRATING THE DETERMINATION OF THE TREND.

Date.	Mean Temperature Fahrenheit in Degrees.	Moving Average, 7 Day Grouping.
Mar. 1	20	
2	25	
3	22	
4	35	24.0
5	26	24.0
6	22	24.4
7	18	26.1
8	20	26.7
9	28	28.7
10	34	29.9
11	39	31.9
12	40	32.7
13	30	33.6
14	32	34.9
15	26	36.1
16	34	37.1
17	43	38.4
18	48	38.9
19	47	41.1
20	39	43.3
21	35	44.3
22	42	
23	49	
24	50	

extreme. Thus, in the above table, 50 might be added at the close three successive times, forming the required new groups in this fashion. Either of these methods is purely an approximation.

It will also be noticed that, in the moving average line shown in Fig. 17, all irregularities have disappeared

THE MOVING AVERAGE OR TREND.

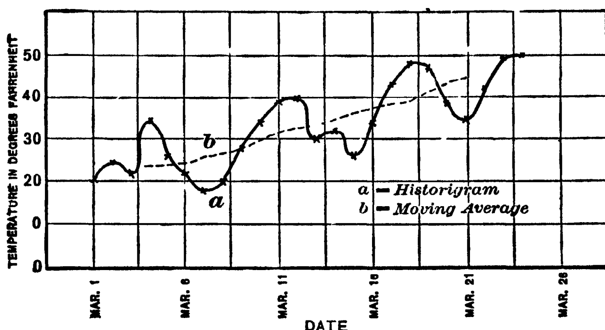


FIG. 17.

and we have obtained the **general rising trend** of temperature during the entire period. **To study short time changes, therefore, the original historigram and not the trend must be studied.**

It is impossible to apply the moving average with equal success to any and all historigrams. Fig. 18 shows a curve in which no regular periodicity is manifested. If a moving average were used in this case, the only possibility would be to take a long period of

perhaps thirty years and this would give nothing but a general trend for the whole time covered without regard to any of the large oscillations which it might be desirable to retain.

Care should be taken in selecting an appropriate vertical scale for all historigrams. If the units occupy too much space, small changes in the size of items will apparently be important fluctuations while, if the units

HISTORIGRAM SHOWING NO DEFINITE PERIODICITY.

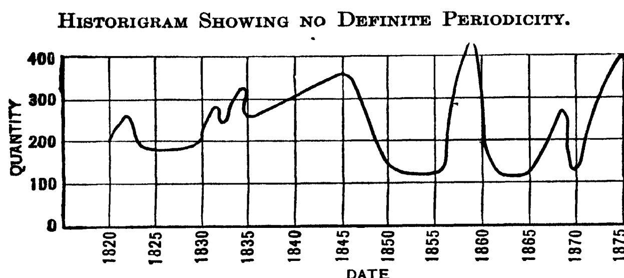


FIG. 18.

occupy too little space, the graph will assume a false appearance of uniformity. This is illustrated by the changed appearance of the wheat-price historigram in Fig. 21 when its vertical scale is relatively much increased by converting it to an index curve.

Sec. 98. **Relative or Proportional Change.**

If the population record of a given city is as follows,

1890	100,000
1900	150,000
1908	200,000

it is evident that the **absolute** increase in the two periods involved was identical, or 50,000 in each case. The **proportional** increase, however, differed for, in the first case, the population had increased by 50 per cent. while, in the latter period, the increase was but 33 per cent. Still, the increase of the first period occurred at the **rate** of 5,000 per year, while, in the second period, **the rate of increase** was larger, being 6,250 annually. The latter, however, is based on a larger population. If we take the base at the beginning of the period we find the **proportional rate of increase** to be, in the first period, 5,000/100,000 or 5 per cent. and, in the second period, 6,250/150,000 or 4.17 per cent. It is the last named quantity, the **proportional rate of change**, in which we are most commonly interested.

To say that the population of New York City increased a million in the last decade and only a hundred thousand in a decade sixty years ago does not give us any idea of the relative change going on for the two periods. In order to remedy this defect, Professor Alfred Marshall has devised a graphic method for readily comparing, on a historigram, the proportional rates of change for different periods. This is illustrated in Fig. 19.

XY is a historigram showing the population of a city at periods ranging from 1840 to the present. It is desired to know whether the proportional rate of growth was greater between 1845 and 1860 or between 1900 and 1910. Let $M'R'$, $D'N'$, MR, and DN be

the respective ordinates for the given years, intersecting the historigram XY at C', A', C and A respectively. Draw $C'B'$ and CB parallel to the base.

HISTORIGRAM ILLUSTRATING MARSHALL'S METHOD OF DEPICTING PROPORTIONAL RATE OF INCREASE.

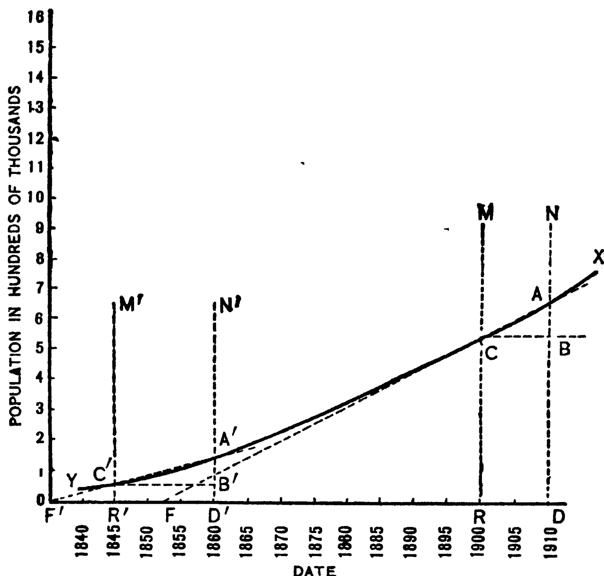


FIG. 19.

Draw $A'C'$ and AC and produce each until they cut the base at F' and F respectively.

Now, $A'B'$ and AB represent the **absolute** increases for the given periods and $A'B'/B'C'$ and AB/BC are

the **rates** of increase for the same. The **proportional rates** of increase are represented by $\dfrac{\frac{A'B'}{B'C'}}{C'R'}$ and $\dfrac{\frac{AB}{BC}}{CR}$ respectively. But $AB/BC = CR/FR$ (corresponding sides of similar triangles). Likewise

$$\frac{A'B'}{B'C'} = \frac{C'R'}{F'R'}.$$

$$\therefore \frac{\frac{AB}{BC}}{CR} = \frac{\frac{CR}{FR}}{CR} = \frac{1}{FR}$$

= the proportional rate of increase 1900–1910.

And

$$\frac{\frac{A'B'}{B'C'}}{C'R'} = \frac{\frac{C'R'}{F'R'}}{C'R'} = \frac{1}{F'R'}$$

= the proportional rate of increase 1845–1860.

Therefore, the proportional **rate** of increase varies directly as $1/FR$ or inversely as FR. In the figure, FR equals approximately $5F'R'$, hence the proportional rate of increase for the period 1900–1910 was only about one fifth of that for 1845–1860.

The merit of the above method lies in its extreme simplicity of application, the only work necessary being to draw the lines AF and CR and measure the lines FR for each period.

Sec. 99. **Logarithmic Historigrams.**

The logarithmic historigram has been devised for the purpose of showing directly on a graph the **pro-**

portional change for all parts of the period considered. It depends on the fact that an equal increase in the logarithm of a number indicates **multiplication** by an equal number and hence an equal proportional change. Fig. 20 shows such a historigram. The proportional increase in population from 1840 to 1845 is indicated by the line AC, that between 1865 and 1875 by the line EF, and that between 1900 and 1910 by the line HI. Since AC approximately equals EF, the **proportional change** for the first and second periods is about the same, but this is not true of the **proportional rate of change.** The latter period was twice as long as the former, hence, the **proportional rate** in the latter case was only about half as great. In the last period 1900–10, the proportional change and also the proportional rate of change is much less.

The proportional **rate** of change is indicated directly by the steepness of the curve at the given point. It may be calculated, approximately, for any period by dividing the altitude of the triangle by the base as AC/BC, EF/DF, etc. The results thus obtained are not exactly correct, and the explanation of this fact brings out one of the weaknesses of the logarithmic curve. When a logarithm is doubled, it does not follow that the base of the logarithm is doubled. Thus, if $AC = 2HI$, it is not true that the proportional increase between 1840 and 1845 is exactly twice that between 1900 and 1910. A quantity when tripled

increases its logarithm by the log. of 3 or 0.477 but, when a quantity is multiplied by six, it does not increase its logarithm by 2×0.477, or 0.954, but by 0.778 instead. Doubling the proportional change then less than doubles the vertical movement of the logarithmic historigram. Bowley[1] seeks to remedy this difficulty by affixing to the diagram a series of vertical lines representing the logarithms of 2, 3, 4, etc. The length of these lines may thus be compared with the vertical change in the logarithmic curve for a given period and

LOGARITHMIC HISTORIGRAM.

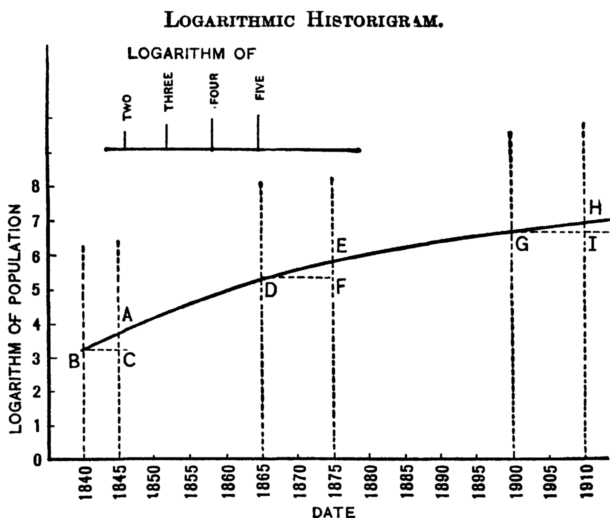


FIG. 20.

[1] *Elements of Statistics*, p. 190.

one may in this manner determine whether population has doubled, trebled, or quadrupled. Such a scale is attached to Fig. 20.

The logarithmic historigram, while valuable for relative comparison in point of time, is not good for comparison of the **sizes** of different variables at the same time. It is of so little value for this purpose that the use of a base line is frequently dispensed with and the several curves shifted vertically until they are in the best position for the easy comparison of the **proportional changes** in each. Thus, if we wish to compare the trend of prices of lumber and steel for a decade we care nothing concerning the relative prices per unit of the two articles but we do wish to know the relative changes in the price of each. Logarithmic curves will show this nicely if placed close to each other or given a common starting point by means of vertical shifting of the whole curves. One disadvantage of logarithmic graphs is that the ordinary reader is unfamiliar with them and unable to correctly interpret their meaning, since it takes practice to get a firm grasp of the idea of relativity contained therein. As a result, it seems best to confine their use for the present primarily to scientific works rather than to utilize them in more popular literature.

Sec. 100. **Index Numbers—General Characteristics.**

Tables of historical statistics are reduced to index numbers for two reasons—first, to facilitate comparison

of the relative changes in two or more synchronous variables; second, to permit of the computation of an average index series.

The sizes of the fluctuations **relative to their respective norms** in a number of simple historigrams cannot be readily compared, especially when the graphs have separate origins or when the quantities are of very different average size. The following hypothetical table and the historigrams shown in the first part of Fig. 21 illustrates this difficulty. When the **absolute** prices are plotted as historigrams, it seems that steel has fluctuated considerably, while wheat has remained almost constant in price. This is, in reality, far from true, the deceptive appearance being due wholly to the comparative units of each variable chosen. Had we taken thirty bushels of wheat instead of one bushel as the basis of price, we should have found the fluctuations more closely allied.

With the given historigrams, a change of 1 mm. on the vertical scale means a large relative variation in the price of wheat, but a very small one in the price of steel. To overcome these difficulties and reduce the two sets of prices to a comparable form, it is best to convert each to an index series. This may be done either by dividing each item of the price series by the price for some years arbitrarily chosen as a base or by dividing each by an average of the whole group. The former is the more common method, but the use of

TABLE XV.

DERIVATION OF INDEX NUMBERS.

Year.	Price of Steel per Ton.	Price of Wheat per Bushel.	Index Price of Steel.	Index Price of Wheat.
1890	$30	$1.05	120	117
1891	27	.96	108	107
1892	24	.94	96	104
1893	22	.83	88	92
1894	24	.88	96	98
1895	26	.92	104	102
1896	22	.72	88	80
	Av. $25	Av. $0.90	Av. 100	Av. 100

HISTORIGRAMS SHOWING PRICE-CHANGES FOR WHEAT AND STEEL.

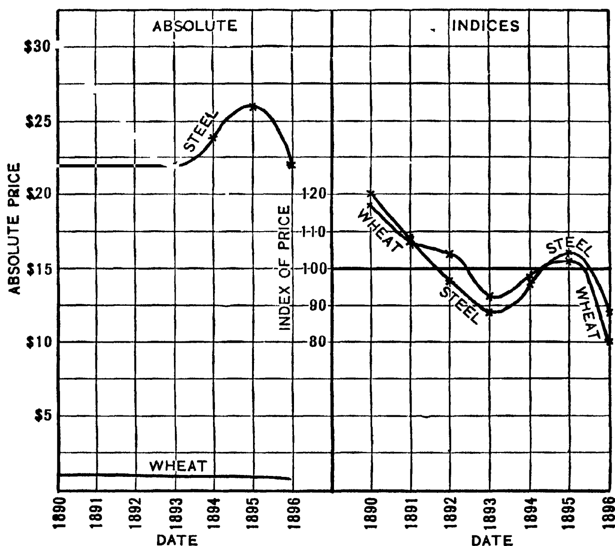


FIG. 21.

the average of a considerable group as a base is more satisfactory, since it is representative and less affected by chance variations. The average of the whole group is the best base of all and the one most generally applicable in statistics. This is the method used in obtaining the indices in the table given above.

When the price indices are plotted instead of the prices themselves, it becomes evident that the changes in the figures for wheat and steel are, in fact, very similar. Now, a vertical fluctuation of 1 mm. in either graph has been made to represent exactly the same **proportional** change in price and the two curves may be legitimately compared.

We see, then, that the reduction of a group of historical data to an index series greatly facilitates the comparison of different synchronous variables with each other, but the index series is no improvement on the original if it is desired to compare different periods in the same series as to the relative changes therein. Index numbers, then, aid in comparisons of the fluctuations of different variables at certain specific dates, but the function of bringing out well the **relative** changes over periods of time is reserved for logarithmic historigrams.

Sec. 101. **Average Indices.**

For many purposes, it is extremely desirable to find the general trend of a large number of variables considered jointly and, to accomplish this, it is necessary

to obtain an average index for each recorded date. Good examples of such average indices are those of prices and wages which are prepared annually by the United States Bureau of Labor. An average price index must be computed from the prices of a large number of articles of varying importance. The first step in the process is to obtain an index series for each article for the entire period, using a common time-base for all the articles. The base used by the United States Bureau of Labor is the decade 1890–9 inclusive. The following table will represent, in miniature, one method of obtaining the average indices for any given dates from the indices of the separate commodities. Several hundred articles might be used in practice instead of the seven cited below.

TABLE XVI.

THE MEDIAN OF INDICES

Date.	Price Indices.							Median Index.
	Wheat.	Cotton.	Steel.	Lumber.	Corn.	Wool.	Leather	
1880	101	120	104	108	103	92	104	104
1881	97	90	102	103	97	99	102	99
1882	95	82	94	100	96	110	96	96
1883	102	108	99	90	100	100	98	100
1884	105	100	101	99	104	99	100	100

The question at once arises as to which kind of average will produce the best results. This depends on the specific nature of the problem. If one wishes to study

the effect of a changing volume of gold or of money on prices, a different average is desirable than the one used if the relative cost of living is the subject of investigation. In the first case, a change in the price of one article is just as good a criterion as a change in price of any other. The quantity or importance of the commodity does not enter into the question at all. If no other factors than the quantity of money were affecting prices every article should fluctuate exactly alike; therefore, the presence of extreme variations from the normal, as of cotton in 1880 or the wool in 1882, are prima facie evidence of extraneous influences which should be excluded in computing the general trend of prices. The average which best eliminates these extreme variations is the median, since it takes no account of them. If we array all the indices for 1880 in order of size, we find the median item to be 104, which is the index for that year. For 1881, the median is 99, and so on to the end of the series. Having obtained an average index for each date, the **general trend of** prices may be clearly shown.

If, however, we desire to know whether the **cost of living** has changed during a given period, an entirely different average is necessary. The average consumer is not recompensed for the fact that the price of meat has gone up by knowing that pepper has fallen an equal extent. Therefore, in computing **a consumers' index,** it is necessary to use an average which ranks each article

according to the amount consumed, and, for this purpose the weighted arithmetic average is best. The weights would not be identical under all conditions of consumption, but would be regulated by the budgets of the group of consumers under consideration. The index of each article is multiplied by a weight corresponding to the percentage of the income spent for this specific commodity. The sum of the products is divided by the sum of the weights to obtain the average consumers' index. This is illustrated in the following table.

TABLE XVII.

WEIGHTED AVERAGE OF INDICES.

Wts.	40		16		14		6		24		Consumers' Index.
Article.	Food.		Rent.		Clothing.		Fuel and Light.		Miscellaneous.		
Date.	Index	Product.	Index	Product	Index.	Product.	Index.	Product.	Index.	Product.	
1891	108	4,320	102	1,632	110	1,540	96	576	104	2,496	106
1892	99	3,960	100	1,600	101	1,414	101	606	101	2,424	100
1893	93	3,720	98	1,586	89	1,246	103	618	95	2,280	94

All decimals omitted in above table.

In 1891, the sum of the products is 10,564 and the sum of the weights is 100, hence the weighted average is approximately 106, this being the consumers' index for that year.

Frequently the original index itself must be an average of minor indices. Thus the food index in the table above is found by obtaining a weighted average

of the indices of each specific food item. This process is used instead of a direct weighted average of the minor items because of the fact that it is seldom possible to get complete figures for every commodity for every year. By using the above method, such gaps are covered without disarranging the whole system.

REFERENCES.

MAYO-SMITH, RICHMOND. *Statistics and Economics.* Macmillan Company, N. Y., 1896, pp. 196–233.

BOWLEY, A. L. *Elements of Statistics*, Chap. VII.

FISHER, IRVING. *The Purchasing Power of Money.* Macmillan Co., N. Y., 1911. Chap. X.

MEITZEN, AUGUST. *Statistics*, pp. 195–204.

BERTILLON, JACQUES. *Cours Elementaire de Statistique*. Chap. XI.

EDGEWORTH, F. Y. *Index Numbers.* Palgrave's Dictionary of Pol. Econ.

ADAMS, THOS. SEWALL. *Index Numbers and the Standard of Value.* Journal of Pol. Econ., Dec., 1901, and Jan., 1902.

FOUNTAIN, H. *The Construction of Index Numbers of Prices.* Board of Trade Report on Wholesale and Retail Prices in the United Kingdom, 1903.

PART IV.

COMPARISON OF VARIABLES.

CHAPTER XVI.

VARIOUS METHODS OF COMPARISON.

Sec. 102. **Purpose and Value of Comparison.**

As was mentioned in the early part of the book, comparison is, in general, the final goal toward which all statistical studies tend. Comparison is necessary to give us clear ideas of the relationship of things in time and space. It is also essential in determining whether phenomena are connected or independent and in establishing relations of cause and effect.

We may wish to study:

1. Changes of a single variable.
2. The structure of different groups.
3. Changes in two or more variables.

We have already discussed in the last chapter by means of historical tables, and historigrams, either simple or logarithmic, the question of changes of a single variable. Most of the methods of comparing the structure of two different groups of data have also been dwelt upon at some length, but perhaps a brief summary of this second case may be helpful.

Sec. 103. **Comparison of Frequency Distribution of Two or More Groups of Data.**

1. **By simple frequency tables and histograms.**

This method is desirable when the aim is to bring out comparisons which show the absolute as well as the relative size of the various classes in the groups compared. It would show, for example, the relative number of men of each grade employed in several different establishents as well as the wage distribution in each place.

2. **By percentage frequency tables and histograms.**

These tables and graphs show nothing concerning the actual size of each group or the classes therein but are vastly superior in making clear the relative distribution between the higher and lower groups in each place. These would not show the actual number of men employed in various establishments, but would bring out distinctly the relative wages paid in each.

3. **Absolute cumulative tables and ogives.**

These are used primarily for the ascertainment of the median, quartiles, deciles, etc., but may be utilized as a substitute for the simple frequency tables and histograms.

4. **Percentage cumulative tables and ogives.**

These are far better than their absolute counterparts for purposes of comparison but are not so good for computing the median, etc.

5. **Lorenz tables and curves.**

For the purpose of showing distribution of wealth, income, etc., at different periods or in different places, these are unequalled.

6. **By coefficients of dispersion.**

These furnish a **numerical** measurement of deviation from the type, a feature lacking in all the previous methods.

7. **By coefficients of skewness.**

By their aid, a **numerical** measurement of the lack of symmetry or the concentration of the items nearer to one than to the other extremity of the group may be shown.

8. **By coefficients of correlation.**

The discussion of these is reserved for a later chapter.

Sec. 104. **Methods of Comparing Changes in Two or More Different Variables.**

Most of the methods of comparison have been discussed briefly and will be merely summarized here. The principal ones are:

1. **By absolute historical tables or historigrams.**

This is the best method of showing the actual changes in different variables. If the wheat crops of the leading nations are thus plotted, both the change in production for each nation during the period and the relative product of each nation at any given time is revealed at a glance.

2. **By index numbers and historigrams.**

These are used when the only desideratum is to compare **changes** and not the **absolute size** of the quantities in the two series. All the curves being reduced to like bases, it is easy to compare the proportional changes relative to the base, in the **different** variables during the **same** period. Thus, we can see at a glance whether the proportional increase in the population of New York from 1900 to 1910 was greater or less than that of Wyoming for the same period. The fact that the absolute increase in New York was vastly larger than that in Wyoming in no way obscures the record of the comparative proportional change. It must be reiterated that the index curves, however, do not indicate the **proportional** change in either state for the last decade as compared with the change of some past decade.

3. **By logarithmic index historigrams.**

Simple logarithmic historigrams show by their vertical movements the comparative **proportional** changes, relative to the preceding period, in two or more variables during the same time-interval. Their respective inclinations from the horizontal at the time of crossing any given time-ordinate indicate the **proportional rates** of change in the different variables at this date. As has been before stated, comparison of logarithmic historigrams is facilitated by vertical shifting of the curves until they are in proximity to each other, the eye being

thus enabled to better follow and compare their trends. This effect is accomplished mathematically instead of mechanically if the original data, in each instance, are reduced to index series and the logarithms of the indices instead of those of the original numbers plotted. Since logarithmic curves are of no practical value for showing the absolute size of the different variables at any given date, nothing whatever is lost by the preliminary reduction to index-series or the consequent vertical shifting of the graphs.

4. **By coefficients of correlation.**

These will be explained later.

Sec. 105. **The Plotting of Comparative Graphs.**

If two or more graphs are to be compared, it is desirable that they be plotted upon the same sheet using the same axes and scales. The use of different colors of ink is one of the best methods of avoiding confusion. When, however, the graphs are to be printed, it is better to adopt different devices for each graph, such as making one line heavier than the other or utilizing dots, dashes, or combinations of the two. It is unwise to place a large number of graphs on one sheet, if the lines lie close together, for they become extremely confusing to the eye. When more than five or six are to be compared, it is best to select one ot the group as a basis of comparison and place it on each sheet, using for this line a heavy ruling in order to differentiate

it from the rest. This rule of course applies equally well to either frequency graphs or historigrams.

Sec. 106. **Long- and Short-time Fluctuations.**

It must be understood that the terms "long" or "short" time are purely relative and that the long term for one variable might be an extremely short period for another. Yet, in most historical series, there appear fluctuations of two or more types occurring contemporaneously. If we were to study the marriage-rate for the past century, we should probably find a more or less steady decline throughout the whole period, but with oscillations covering five to ten year periods marking the epochs of prosperity or financial depression and still a third series of annual waves whose crests would be found in the month of June. Each of these three variations is due to a different cause but all three causes are acting simultaneously. Likewise, a study of the weather changes will reveal a cycle of five or six days duration, due to the regular procession of the cyclones across the United States, an annual cycle, due to the passage of the earth around the sun, and a cycle of some fifteen years, due to causes as yet unknown. In order to study any one of these cycles by itself, it is necessary to adopt the procedure of the physicist and eliminate, in so far as possible, all the other factors. Unfortunately, the statistician can rarely indeed, like the physicist, control the conditions of his experiment,

but he can do the next best thing by ridding, so far as possible, the recorded data of the apparent effects of the extraneous causes. If we desire to study the long time changes in unemployment, it will be much better if the seasonal fluctuations can be removed from the field. This is best done by applying the moving average according to the rules given in Sec. 97. The wave-length, here, is evidently one year, hence the groups for the moving average must cover a period of that length.

Sec. 107. **The Elimination of Long-time Variations.** We are frequently interested in the short-time oscillations only and desire, therefore, to eliminate all long-time changes. If, for example, we are studying seasonal fluctuations in unemployment, we find the oscillations due to crises or disturbances of industry serious hindrances in our investigation. A simple way to study strictly seasonal changes is to obtain a seasonal average for a series of years. If monthly records only are available, the process would be as shown in the table on the next page. By this average, the typical trend of unemployment throughout the season is made apparent.

Another very important method of eliminating long-time fluctuations in a single given historigram is as follows. **The data are plotted as a graph, the proper period selected, and the moving average line computed. The deviations of the original data from the trend are**

now found and tabulated. These deviations are finally plotted on a horizontal base line.

TABLE XVII.

PERCENTAGE OF UNEMPLOYMENT.

Month.	Year.						
	1900	1901	1902	1903	1904	1905	Average.
Jan.	4.1	5.0	4.7	12.6	7.4	5.1	6.5
Feb.	3.6	4.8	5.2	12.1	8.1	4.6	6.4
Mar.	3.2	4.1	5.4	9.2	5.2	4.3	5.2
Apr.	2.1	2.7	3.8	7.4	6.1	4.2	4.2

Sec. 108. **Comparison of the Long- or Short-time Fluctuations in Two Historigrams.**

These are two of the commonest and most important of all processes in the field of comparative statistics. The short-time deviations are compared by reducing each of the variables to index numbers, applying the rule just given, and plotting both of the final series of deviations on the same base line. The process is illustrated in the following hypothetical table and the graphs obtained therefrom are shown in Figs. 22 and 23.

Fig. 22 shows the elimination of the short-time changes by means of the moving average. Graph *d* shows the steady increase of the **supply** of the given commodity up to 1904 when the production begins to fall off slightly. Graph *c* pictures the trend in **price.** Until 1894, this falls as the supply increases just as would normally be expected but, after that date, the

TABLE XVIII.

COMPARISON OF SHORT-TIME FLUCTUATIONS ONLY IN THE SUPPLY AND PRICE OF A COMMODITY.

Date.	Supply. Index of Supply	Supply. Moving Average of Indices.	Supply. Deviations from Moving Average.	Price. Index of Price.	Price. Moving Average of Indices.	Price. Deviations from Moving Average.
1880	80			146		
1	82			140		
2	86	84	+2	130	133	− 3
3	91	85	+6	117	129	−12
4	83	87	−4	133	124	+ 9
5	85	89	−4	127	117	+10
6	89	89	0	115	114	+ 1
7	96	91	+5	95	109	−14
8	93	92	+1	100	104	− 4
9	90	93	−3	106	100	+ 6
1890	91	94	−3	103	96	+ 7
1	94	96	−2	94	89	+ 5
2	100	98	+2	75	83	− 8
3	105	99	+6	66	80	−14
4	102	100	+2	75	79	− 4
5	96	101	−5	91	80	+11
6	98	103	−5	87	82	+ 5
7	106	105	+1	81	83	− 2
8	114	108	+6	76	83	− 7
9	112	109	+3	82	86	− 4
1900	109	111	−2	91	88	+ 3
1	106	112	−6	100	88	+12
2	112	113	−1	89	88	+ 1
3	120	114	+6	76	89	−13
4	118	114	+4	82	91	− 9
5	112	113	−1	100	96	+ 4
6	110	112	−2	106	101	+ 5
7	107			114		
8	113			103		
	Av. 100			Av. 100		

Note.—The number in the above table are accurate to units place, all decimals being dropped.

price rises steadily despite the increasing supply. This may be due to a falling off of the supply in other countries or to an increase in the demand for the com-

INDEX HISTORIGRAMS WITH MOVING AVERAGES INDICATING THE RELATIONSHIP OF SUPPLY AND PRICE.

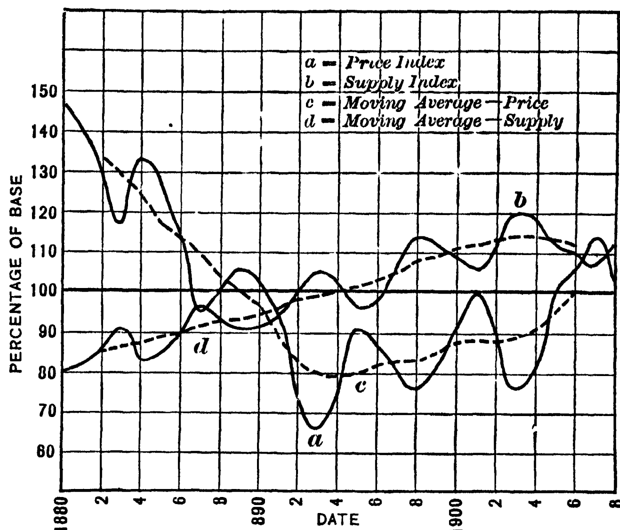


FIG. 22.

modity, due perhaps to an increase in population or prosperity or perhaps to a change in the habits of the people.

In Fig. 23, the process is reversed and the long-time oscillations are eliminated. Just imagine that each of the moving average lines in Fig. 22 is stretched out

straight and then superimposed on the axis MN in Fig. 23, without detaching from the moving average lines the original historigrams from which the trends were derived. This, in effect, is what has been done in Fig. 23. When placed in this position, we perceive a marked relationship between the two curves, now

SHORT-TIME OSCILLATIONS — SUPPLY AND PRICE.

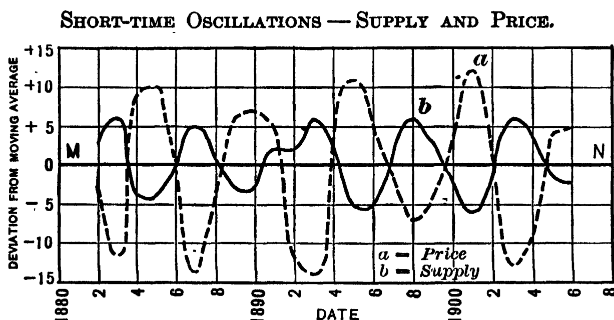


FIG. 23.

totally unobscured by the long-time changes. When a rises, b falls and vice versa. Since the comparison of short-time oscillations is needed more frequently than that of long-time changes, this method is of great practical importance.

REFERENCES.

YULE, G. U. *Introduction to Statistics*, Chaps. X and XI.

BOWLEY, A. L. *Elements of Statistics*, Chap. VII.

HOOKER, R. H. *Correlation of Successive Observations.* Jour. of the Royal Statistical Society, LXVIII, 696–703.

NORTON, J. P. *Statistical Studies in the N. Y. Money Market.* Macmillan and Co., N. Y., 1902.

CHAPTER XVII.

CORRELATION.

Sec. 109. Definition of Correlation.

In the last chapter we have studied various methods of comparing data and, when we study comparison, we closely approach the subject of correlation. **Correlation means that between two series or groups of data there exists some causal connection.** Investigation might prove that the cocoanut crop of the Fiji islands had been increasing steadily while the money supply of the United States had done likewise. This would not imply correlation unless it could be shown that one was the cause of the other or that both changes were due to some common third factor. It, therefore, is often approximately true that $a \propto b$ when there is no correlation but, if correlation does exist, then, necessarily, the changes in a must bear some fixed relationship to the changes in b.

Sec. 110. Kinds of Correlation.

If extraneous factors could be absolutely eliminated, it would then follow that a would, in every case, bear an exact mathematical relationship to b, which might be for example $a \propto b$, $a \propto 1/b$, $a \propto \sqrt{b}$, $a \propto (b + x)$, etc. Since, in practice, the influences acting are so numerous and complex it is impossible to rid the statistics of all

that are undesirable, usually but one or two of the most important being thus removed from the field. As the minor inharmonious factors still remain to confuse the results, it is seldom, especially in the field of social statistics, that any absolutely fixed mathematical relationship between two variables can be established. We very often must be satisfied if we learn that when one variable increases there is a certain tendency for the other to increase or vice versa. **If it is proven true that in a large number of instances two variables tend always to fluctuate in the same or in opposite directions we consider that the fact is established that a relationship exists. This relationship is called correlation.** If the two graphs constantly deviate in the **same direction** at the same time we say that there is **direct** correlation, while if they steadily deviate in **opposite directions** we call it **inverse** correlation. In Fig. 21 we see that the price of wheat and the price of steel tend to oscillate regularly in similar directions, therefore, the correlation is direct, but in Fig. 23, on the contrary, the price curve always rises when the supply curve falls, hence, as far as the short-time relationship is concerned, the correlation is inverse. Since, in both of these instances, the changes of the two variables are closely related throughout the entire period, we would say that there was a **high degree** of correlation. In the first instance, the price of steel and the price of wheat are presumably both dependent on some other

factor, such as a variation in the volume of money or general business conditions, while, in the second case, the change in the supply is the direct cause of the variation in price.

Sec. 111. **Correlation Applied.**

We may study the correlation between two historical variables or between any other two groups of related phenomena. Examples of the first would be the correlation of the gold movement with exports, the marriage rate with the price of wheat, the amount of unemployment with the bank clearings, etc. The second might be illustrated by the relation between the length and breadth of leaves. As a leaf grows longer does it always grow broader as well or does the breadth tend to remain constant? Do tall fathers always have sons taller than the normal? If so, there is correlation. Correlation is often studied by means of index historigrams from which the undesirable oscillations have been, as far as possible, eliminated. Correlation in two variables may be roughly illustrated in frequency graphs. Graphic methods, however, are all somewhat deficient since they cannot give a numerical measurement of the **degree** of correlation existing. For this purpose, we must compute a correlation coefficient, in other words a numerical measurement of the degree to which correlation exists between the **subject** and the **relative** as the two variables to be compared are called. By the term "subject" we mean the variable which is

to be used as a standard or measure and, by the term "relative" we designate the variable which is to be compared with or measured in terms of the subject.

A coefficient of + 1 indicates perfect direct correlation, one of 0 indicates no correlation whatever, and a coefficient of − 1 means that the correlation is perfect but inverse.

Sec. 112. **Karl Pearson's Coefficient of Correlation.** When two different characteristics of a given series of items, as, for example, the length and breadth of leaves, or a given characteristic in related pairs of items, such as the stature of fathers and sons, the ages of husband and wife, etc., are to be compared, or when we wish to study the relationship between the **long-time** changes in two historical variables, the most satisfactory coefficient of correlation is that devised by the great biologist Karl Pearson.

Let x_1, x_2, x_3, etc., be the deviations of the items of the subject from the arithmetic average and y_1, y_2, y_3, etc., be the corresponding deviations of the items in the relative. Let σ_1 be the standard deviation of the subject and σ_2 be the standard deviation of the relative. Let n equal the total number of pairs of items. Let r represent Karl Pearson's coefficient[1] of correlation.

[1] For the derivation of the formula for this coefficient see G. Udny Yule's "Introduction to the Theory of Statistics," pp. 168–174.

Then

$$r = \frac{\Sigma(xy)}{n\sigma_1\sigma_2}.$$

TABLE XIX.

COMPUTATION OF KARL PEARSON'S COEFFICIENT OF CORRELATION FOR AGES OF HUSBAND AND WIFE.

Subject.			Relative.			
h Age of Husband.	x Deviation of Age from Average.	x^2	w Age of Wife	y Deviation of Age from Average.	y^2	xy
22	−8	64	18	−8	64	+64
24	−6	36	20	−6	36	+36
26	−4	16	20	−6	36	+24
26	−4	16	24	−2	4	+ 8
27	−3	9	22	−4	16	+12
27	−3	9	24	−2	4	+ 6
28	−2	4	27	+1	1	− 2
28	−2	4	24	−2	4	+ 4
29	−1	1	21	−5	25	+ 5
30	0		25	−1	1	0
30	0		29	+3	9	0
30	0		32	+6	36	0
31	+1	1	27	+1	1	+ 1
32	+2	4	27	+1	1	+ 2
33	+3	9	30	+4	16	+12
34	+4	16	27	+1	1	+ 4
35	+5	25	30	+4	16	+20
35	+5	25	31	+5	25	+25
36	+6	36	30	+4	16	+24
37	+7	49	32	+6	36	+42
Av. 30	$\Sigma x^2 = 324$		Av. 26		$\Sigma y^2 = 348$	$\Sigma(xy) = +287$

The mode of computing this coefficient is shown in

the following table in which the respective ages of husband and wife are placed on the same line. This must be considered as a representative group of sample data standing for a very large number of items.

$$r = \frac{\Sigma(xy)}{n\sigma_1\sigma_2} = \frac{287}{20 \times 4.02 \times 4.17} = \frac{287}{335.27} = +.856.$$

A study of this table will make plain that the numerator principally regulates the size of the coefficient. If, in any pair of items, both subject and relative are larger or both smaller than their respective averages, then *xy* will be positive but, if the subject is larger than the average while the relative is smaller, then *xy* is negative. But many negative values for *xy* will make the coefficient small while if the values are nearly all positive it is likely to be quite large. If *xy* is quite uniformly negative, then the coefficient will represent a high degree of inverse correlation. It is evident, then, that the comparative location of the items of the respective pairs with regard to the average is the important criterion in this coefficient. The distance from the average is, relatively, of less moment.

Sec. 113. **The Application of Karl Pearson's Coefficient to Long-time Changes in Historical Variables.**

This coefficient may be applied as well to historical data as to the variations in items at any specific time. In its **original form,** however, this coefficient can only

be used, with the former, in connection with the **long-time** changes. In Fig. 22, for instance, curves *a* and *b* both tend to remain on opposite sides of the arithmetic average line, hence a negative coefficient would be obtained. The size of this coefficient would be affected but slightly by the short-time oscillations, since few of these cross the average line at all. The correlation between the short-time oscillations is seen at a glance to be inverse.

INDEX HISTORIGRAMS.

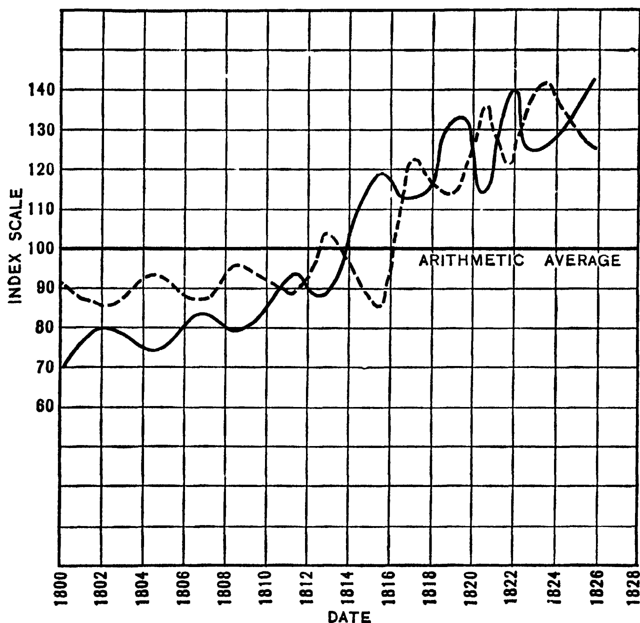


FIG. 24.

TABLE XX.

ILLUSTRATING THE CORRELATION OF SHORT-TIME OSCILLATIONS BY MEANS OF PEARSON'S MODIFIED COEFFICIENT.

Date.	Supply.				Price.				xy
	Index of Supply.	Moving Average of Indices.	x Deviations from Moving Average.	x^2	Index of Price.	Moving Average of Indices.	y Deviations from Moving Average.	y^2	
1880	80				146				
1	82				140				
2	86	84	+2	4	130	133	− 3	9	− 6
3	91	85	+6	36	117	129	−12	144	−72
4	83	87	−4	16	133	124	+ 9	81	−36
5	85	89	−4	16	127	117	+10	100	−40
6	89	89	0	0	115	114	+ 1	1	0
7	96	91	+5	25	95	109	−14	196	−70
8	93	92	+1	1	100	104	− 4	16	− 4
9	90	93	−3	9	106	100	+ 6	36	−18
1890	91	94	−3	9	103	96	+ 7	49	−21
1	94	96	−2	4	94	89	+ 5	25	−10
2	100	98	+2	4	75	83	− 8	64	−16
3	105	99	+6	36	66	80	−14	196	−84
4	102	100	+2	4	75	79	− 4	16	− 8

TABLE XX—*Continued.*

Date.	Supply.				Price.				xy
	Index of Supply.	Moving Average of Indices.	x Deviations from Moving Average.	x^2	Index of Price.	Moving Average of Indices.	y Deviations from Moving Average.	y^2	
1895	96	101	−5	25	91	80	+11	121	− 55
6	98	103	−5	25	87	82	+ 5	25	− 25
7	106	105	+1	1	81	83	− 2	4	− 2
8	114	108	+6	36	76	83	− 7	49	− 42
9	112	109	+3	9	82	86	− 4	16	− 12
1900	109	111	−2	4	91	88	+ 3	9	− 6
1	106	112	−6	36	100	88	+12	144	− 72
2	112	113	−1	1	89	88	+ 1	1	− 1
3	120	114	+6	36	76	89	−13	169	− 78
4	118	114	+4	16	82	91	− 9	81	− 36
5	112	113	−1	1	100	96	+ 4	16	− 4
6	110	112	−2	4	106	101	+ 5	25	− 10
7	107				114				
8	113				103				
	Av. 100		Total	358	Av. 100		Totals	1,593	−728

Note. — The figures in the above table are only read to the nearest unit.

Fig. 24 shows us a case in which the long-time changes are in the same direction while the short-time oscillations are in opposite directions. By the use of Pearson's method, we would obtain a large positive coefficient, but this would take no account whatever of the inverse relationship existing between the short-time fluctuations. In computing the coefficient for the long-time changes, the method used in Sec. 112 is followed throughout, the items and deviations for the same date being paired together.

Sec. 114. **The Modification of Karl Pearson's Coefficient for Use with Short-time Oscillations.**

The modification of the historigrams in order to bring out distinctly the short-time fluctuations suggests that a like process will be useful in obtaining a coefficient of correlation and this is the method actually followed. In Fig. 23 the long-time changes have been completely eliminated. The deviations of the items from the **trend** instead of from the **arithmetic average** are now taken into consideration in computing the x and y columns. These deviations are also squared and used in computing σ_1 and σ_2. Table XVIII, when completed for this purpose, appears as shown in Table XX. Using the results obtained from the latter table we find

$$\sigma_1 = \sqrt{\frac{\Sigma x^2}{n}} = \sqrt{\frac{358}{25}} = \sqrt{14.32} = 3.78,$$

$$\sigma_2 = \sqrt{\frac{\Sigma y^2}{n}} = \sqrt{\frac{1,593}{25}} = \sqrt{63.72} = 7.98.$$

$$r = \frac{\Sigma(xy)}{n\sigma_1\sigma_2} = \frac{-728}{25 \times 3.78 \times 7.98} = -.965.$$

It will be observed that, in the foregoing table, $n = 25$, since only the years from 1882 to 1906 inclusive can be used in computing the coefficient. The indices, moving averages, and deviations are carried only to units place. Greater precision might be obtained by carrying out the decimals. The coefficient of $-.965$ shows a very high degree of inverse correlation between supply and price, indicating that an increase in supply means almost invariably a fall in price and vice versa.

Sec. 115. **The Coefficient of Concurrent Deviations.**

Karl Pearson's coefficient, when modified as described above, is an excellent method of comparing short-time fluctuations, but it requires considerable time and effort for its computation. There is another method of obtaining a coefficient of correlation which has the merit of extreme simplicity and, in most cases, may be used satisfactorily in the study of short-time oscillations. It is wholly worthless for dealing with long-time changes, since it takes almost no account of the general trend.

If, in the comparison of two historigrams, it is noticed that both curves tend to move in the same direction at the same time—that is, if the deviations are concurrent—we say that there is evident direct correlation between the short-time oscillations. If, at each

given date, the curves are moving in opposite directions—the deviations being divergent—we know that the short-time movements are inversely related. In this case, we consider not the deviations from the arithmetic averages or from the moving averages but simply from the item of the last preceding date recorded. The **size** of the deviation is not taken into account but only its **direction.** This fact sometimes constitutes a weakness in this coefficient, for a **trivial** change from the preceding year is often due to some cause other than the principal ones which alone can be taken into consideration, but the slightest change has just as great a weight as the largest if in the same direction. When, however, the pairs of items are numerous, the effects of such chance errors are not likely to prove serious.

The following empirical formula is used to bring the coefficient into a form similar to others, that is, to make + 1 indicate perfect direct correlation, − 1 perfect inverse correlation and 0 no correlation at all.

If r = the coefficient of correlation,
n = the number of pairs of items,
c = the number of concurrent deviations,

then

$$r = \pm \sqrt{\pm \frac{2c - n}{n}}.$$

The use of the signs requires a word of explanation. If the quantity $\frac{2c - n}{n}$ is negative the sign (−) is

TABLE XXI.

CORRELATIONS OF SHORT-TIME FLUCTUATIONS OF SUPPLY AND PRICE BY MEANS OF CONCURRENT DEVIATIONS.

Date.	Supply.		Price.		Product. xy
	Indices of Supply.	x Deviations from Preceding Year.	Indices of Price.	y Deviations from Preceding Year.	
1880	80		146		
1	82	+	140	−	−
2	86	+	130	−	−
3	91	+	117	−	−
4	83	−	133	+	−
5	85	+	127	−	−
6	89	+	115	−	−
7	96	+	95	−	−
8	93	−	100	+	−
9	90	−	106	+	−
1890	91	+	103	−	−
1	94	+	94	−	−
2	100	+	75	−	−
3	105	+	66	−	−
4	102	−	75	+	−
5	96	−	91	+	−
6	98	+	87	−	−
7	106	+	81	−	−
8	114	+	76	−	−
9	112	−	82	+	−
1900	109	−	91	+	−
1	106	−	100	+	−
2	112	+	89	−	−
3	120	+	76	−	−
4	118	−	82	+	−
5	112	−	100	+	−
6	110	−	106	+	−
7	107	−	114	+	−
8	113	+	103	−	−

introduced before it and also before the radical. This is necessary in order that the square root may be extracted and the result retain the same sign as that of the original quantity.

The derivation of this coefficient from the same data used in Table XX is illustrated in Table XXI. In this latter table,

$$n = 28, \quad c = 0.$$

Then

$$r = \pm \sqrt{\pm \frac{2c - n}{n}},$$

$$r = \pm \sqrt{\pm \frac{0 - 28}{28}},$$

$$r = - \sqrt{-(-1)},$$

$$r = - \sqrt{1},$$

$$r = - 1.$$

Therefore the inverse correlation is perfect.

Let us suppose another case in which our study covers 48 periods, n therefore equalling 47. If 16 pairs of deviations were concurrent the formula would appear thus:

$$r = \pm \sqrt{\pm \frac{2c - n}{n}} = \pm \sqrt{\pm \frac{32 - 47}{47}}$$

$$= - \sqrt{- \frac{-15}{47}} = - \sqrt{.3191} = - .56.$$

This would indicate then only a moderate degree of inverse correlation.

In addition to its simplicity, this coefficient has the

merit of being well suited for use with irregular graphs like the one shown in Fig. 18 in which smoothing by means of a moving average is well-nigh impossible.

We have used index numbers as examples in the tables cited but this was only done in order to compare the coefficients with the historigrams shown in the illustrations. **In the computation of either Karl Pearson's coefficient or that of concurrent deviations, nothing is gained by reducing the data to indices** and the process tends to introduce slight mathematical errors.

Sec. 116. **The Use of the Lag.**

We oftentimes find, in the comparison of two historigrams, that, while there is evident correlation, the crests and troughs of the waves do not quite coincide in the two graphs. This may be due to the necessity of an interval of time elapsing between cause and effect. Unemployment naturally gives rise to poverty and poverty to pauperism but it takes a considerable period of time for pauperism to result from the lack of work.

When one feels sure that one given phenomenon is the cause of another and it seems probable that a period of time should intervene between cause and effect, the graph representing the effect should be lagged sufficiently to make the wave crests coincide—that is, the dates coupled together should not be identical. The required length of lag can be best determined by a study of the two historigrams. Fig. 25

illustrates a case in which there is marked correlation but curve *b* lags behind curve *a* about one month. In computing a correlation coefficient, therefore, it would be necessary to pair together the figures for *a* in May and *b* in June, *a* in June with *b* in July, etc.

THE LAG IN HISTORIGRAMS.

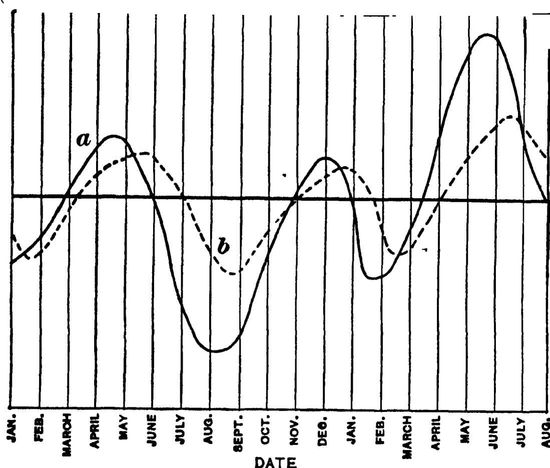


FIG. 25.

One should never assume a certain lag merely from the appearance of the graphs for, in that way, one may easily transpose cause and effect. By lagging curve *b* we obtain a high degree of direct correlation but, if we lagged *a*, we would find just as marked inverse correlation. We must, then, determine from information outside the graphs what the general time relationship

should be and utilize the graphs merely to obtain the proper length of that time interval. The lag should be used, whenever necessary, in the construction of comparative historigrams and also in the computation of coefficients of correlation. An example of the use of the lag is given in Fig. 25 and in Table XXII, accompanying Sec. 121.

Sec. 117. **The Probable Error.**

If we find that two variables fluctuate together in two or three different instances, it by no means follows that this is a proof of the existence of correlation any more than would the fact of throwing double sixes with a pair of dice three times in succession prove that there was any connection between the dice. Such coincidences are likely to be entirely due to chance. If, however, double sixes were thrown ten times in succession we would suspect that some other force than chance was at work; so, when we find an apparent relationship between the fluctuations of two graphs at many different dates, we believe that the chances of an actual relationship existing are very large.

Were we to find 55 pairs of deviations out of 100 concurrent and 45 divergent we would presume that the inequality was due entirely to chance but if 70 pairs were concurrent and only 30 pairs divergent, the probability of this being due to chance alone would be extremely slight.

Thus. we see that the probable error of a coefficient

of correlation varies inversely both with the number of pairs of items and with the size of the coefficient. The law of probable error has been carefully worked out by mathematicians, and the following formula evolved by processes too complex to be given in a text of such an elementary nature as this one.[1]

If r = the coefficient of correlation
and n = the number of pairs of items.
Then

$$\text{the probable error} = \frac{.67(1 - r^2)}{\sqrt{n}}$$

This formula presupposes a purely chance grouping of the data. It deals only with the error arising from the small size of the coefficient or the limited number of items and it is in no way affected by irregularities in the size of the items or the size of classes. It means that the coefficient of correlation should always be written

$$r \pm \frac{.67(1 - r^2)}{\sqrt{n}}$$

and indicates that the chances are that r actually lies between

$$r + \frac{.67(1 - r^2)}{\sqrt{n}} \quad \text{and} \quad r - \frac{.67(1 - r^2)}{\sqrt{n}}$$

[1] For a discussion of this point see Bowley, A. L., *Elements of Statistics*, Part II.

Sec. 118. The Interpretation of the Coefficient of Correlation.

The following rules will assist in giving a general idea of the interpretation of r according to its relation to its probable error:

1. If r is less than the probable error, there is no evidence whatever of correlation.

2. If r is more than six times the size of the probable error, the existence of correlation is a practical certainty.

There might be added to the above the further statements that, in those cases in which the probable error is relatively small.

1. If r is less than .30 the correlation cannot be considered at all marked.

2. If r is above .50 there is decided correlation.

REFERENCES.

ELDERTON, W. P. and E. M. *Primer of Statistics*, Chaps. V and VI.

PEARSON, KARL. *The Grammar of Science.* London, 1900. Pp. 381–92

BOWLEY, A. L. *Elements of Statistics*, Part II, especially pp. 315–334.

YULE, G. U. *Introduction to Statistics*, Chaps. IX–XII and XVI.

BOWLEY, A. L. *Measurement of Groups and Series*, Lectures V and VI.

"STUDENT." *On the Probable Error of a Correlation Coefficient* Biometrika, 1908, VI, 302.

CHAPTER XVIII.

THE RATIO OF VARIATION.

Sec. 119. The Ratio of Variation Defined.

It was stated in Sec. 110 that, if all extraneous influences could be completely eliminated, a given cause would always bear a definite mathematical relationship to its effect. It was also shown that such complete elimination is practically impossible. Nevertheless, it is frequently possible to establish the mathematical relationship actually existing with an approximate degree of accuracy. The possible relationships are numerous, but we shall deal with but one of the most common.

In Chap. XI, the fact was noted that items of a given variety tend to cluster about some specific type or mode. Thus, we have a modal length for leaves, a modal height for men and a normal price for wheat. The mode is usually located quite close to the arithmetic average, the latter quantity being, as a rule, more definite than the former and hence, in many instances, better suited for use as a base, though, in cases in which the mode is well defined, it is probably preferable.

Because two variables constantly fluctuate together, either directly or inversely, it by no means follows that the deviations from the type or mean will be absolutely or proportionally the same. In other words, the wave-

length may be identical but the altitudes markedly different. This fact may be compared to the vibration of two pendulums both of which have the same length and the same period of oscillation, though one swings through an arc of 5° and the other through one of 20°. Fig. 23 gives us an illustration of this case, curve *a*, representing price, swinging approximately twice as far in each direction as does curve *b*, which stands for supply. Such deviations in price would be typical for commodities the demand for which is inelastic. It is frequently desirable to know just what is the average ratio between the proportional or percentage deviation of the two curves from their respective types. The grain speculator is anxious to know how much the price of corn will be raised if the normal crop is 10 per cent. short. The reformer may be desirous to learn in what proportion liquor sales diminish when the number of saloons is halved. The sociologist is interested in finding out to what an extent the birth-rate is affected by the increase in the age of the parents at marriage. This relationship may be entitled **the ratio of variation.** Since it is a corollary of the coefficient of correlation and so closely related thereto, it has frequently been confused with the same but the two are, evidently, radically different in their nature.

Since the proportional fluctuations of supply, illustrated in Fig. 23, are, on the average, about one half as large as the price oscillations, the **ratio of variation,** in that case, would, therefore, be approximately 0.50.

Sec. 120. Computing the Ratio of Variation.

As was indicated in the last section, we desire to find the **average** ratio of the proportional deviations from the type of the items in the relative as compared with those of the subject. The first step is to determine which of the variables shall be taken as the subject and which shall be considered as the relative. In the biological field, it usually makes little difference which series is chosen for each, but, in studying the social sciences, the series having the **larger** average **proportional** deviations is taken as the subject. This is done in order that the ratio may be expressed as less than unity. In Fig. 23, therefore, the price curve *a* would be chosen as the subject and the supply curve *b* as the relative.

One way of computing the **ratio of variation** for historical series would be to take the deviation of the relative from the mean at each given date and divide it by the corresponding deviation of the subject, summate all the quotients thus obtained, and then find the average ratio by dividing by the number of quotients. Thus, if at any given ordinate, the deviation of the subject is 16 per cent. and that of the relative 4 per cent. the quotient would be 0.25, but if the deviation of the subject were − 10 while that of the relative was + 2 the quotient would be − 0.20 and so would tend to reduce the sum. If the oscillations were all perfectly regular, this system of determining the ratio would be satisfactory, but in

practice it cannot be well used, so that other and better methods have been devised, the best of these being the Galton graph devised by Professor Francis Galton.

Sec. 121. **The Galton Graph.**

In applying this graph to two historical variables, it is first necessary to reduce each to an index series. When it is desired to study **long-time** changes **or when no well-defined trend is discernible,** the index numbers are obtained by dividing each item by the **arithmetic average.** The pairs of indices are then plotted, using the subject as ordinate and the relative as abscissa in each case. The method of procedure may best be illustrated by an example.

The following table shows the actual bank clearings and immigration statistics for the United States from 1880 to 1896[1] and the same reduced to index numbers for use in the modified form of Galton graph shown in Fig. 26. The bank clearings for the year are a measure of prosperity. Business conditions, however, affect principally the immigration for the succeeding year, hence it is necessary to introduce a **lag** (see Sec. 116) of one year and compare the bank clearings for 1880 with the immigration for 1881. Since the fluctuations in the immigration index are larger than those of the index for bank clearings, the **former** will be used as the **subject** and the **latter** as the **relative.**

[1] See Statistical Abstract of the United States for 1909, pp. 711 and 753.

TABLE XXII.

DATA FOR GALTON GRAPH TO SHOW THE RATIO OF VARIATION BETWEEN BANK CLEARINGS AND IMMIGRATION.

Subject.			Relative.		
Date.	Immigrants in Tens of Thousands.	Index of Immigrants.	Date.	Bank Clearings in Billions.	Index of Bank Clearings.
1881	67	136	1880	37	106
1882	79	161	1881	49	140
1883	60	122	1882	47	134
1884	52	106	1883	40	114
1885	40	81	1884	34	97
1886	33	67	1885	25	71
1887	49	100	1886	33	94
1888	55	112	1887	35	100
1889	44	90	1888	31	89
1890	46	94	1889	35	100
1891	56	114	1890	38	109
1892	62	126	1891	34	97
1893	50	102	1892	36	103
1894	31	63	1893	34	97
1895	28	57	1894	24	69
1896	34	69	1895	28	80
	Av. 49.1			Av. 35.0	

Sec. 122. **The Ratio of Variation.**

In the Galton graph, it is best to plot the subject on the vertical and the relative on the horizontal scale. When this is done for the preceding data, we obtain a number of points, some rather widely scattered, but tending, in general. to form a band running downward to the left. The next step is to draw the line most nearly approaching the general trend of the dots. This is usually done by finding a line running in the correct

direction, as nearly as can be located by the eye, and having an equal number of points on each side of it.

MODIFIED GALTON GRAPH SHOWING RATIO OF VARIATION BETWEEN BANK CLEARINGS AND IMMIGRATION, UNITED STATES, 1880–1896.

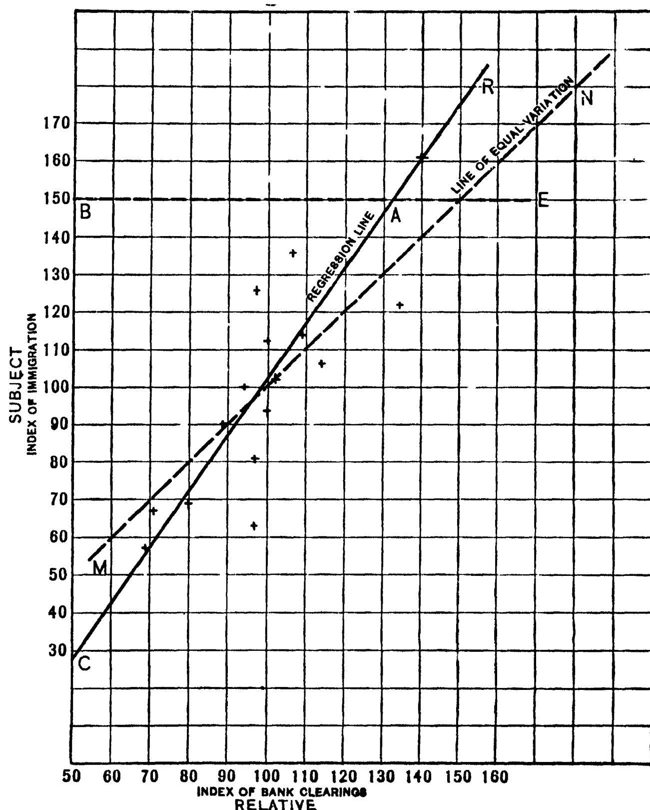


FIG. 26.

The line *RC*, in Fig. 26, approximates this result. **If the correlation were perfect, every point would be located either on a straight line or on some well-defined curve.** While it is probably true that the mathematical relationship existing between subjects and relatives in general are such that, if they could be exactly ascertained, the graphs resulting would more frequently be curves than straight lines, yet it is found, in practice, that the points plotted are usually so widely scattered that a straight line approaches them all about as closely as any mathematical curve that can be found. Hence, in most instances, a straight line and not a curve is drawn since the straight line alone can be used for computation of the ratio of variation according to the simple method devised by Galton. In Fig. 26, the points are so widely scattered that the influence of the external forces in hiding the correlation, as well as the correct ratio of variation, is clearly evident.

The following rules apply to Galton graphs in general: If the line of points slopes downward to the **left,** the correlation is **direct** but if it slopes downward to the **right,** then the correlation is **inverse.** If the points are so badly scattered that they show no definite tendency to form a straight line or regular curve then no correlation is indicated, but the more closely they approach the linear form, the larger the coefficient of correlation to be expected.

If the relative increases or decreases 1 per cent. for

a change of a like amount in the subject, then the ratio of variation is evidently unity. In this case, the points should range themselves along a line of 45° slope such as *MN*, which represents the line of **equal proportional variation.** If, however, the relative shows a tendency to change less, proportionally, than the subject, the angle with the vertical will be less than 45°. In this case, the line along which the points range themselves is called the **regression line.** It receives its name from the fact that, in the biological field, it has been noted that the forces of heredity do not lead offspring to inherit, in full amount, the peculiarities of the parents. If both parents are one inch taller than the normal the chances are that their children will be something **less than one inch above** the average stature, in other words, they **regress** toward the type, hence the term regression. In general, a very large degree of regression, that is, a large deviation of line *RC* from *MN* indicates a slight degree of correlation. When *RC* is nearly vertical, it shows that the relative is but slightly affected by the changes in the subject. If, for example, we were studying the sons of tall fathers and it was found that their stature was but slightly above the normal, there would be a strong probability that the apparent variation was due wholly to chance and that they were really no taller, on the average, than the sons of men of average height.

It must also be remembered that a regression line

based on a few instances is likely to be far from accurate, and, when the probable error of the coefficient of correlation is so large as to vitiate that coefficient, it is practically useless to attempt to compute a ratio of variation.

If BE is any horizontal line cutting the line of regression at A, then, manifestly, the ratio of the average variations of the relative to the average variation of the subject is represented by AB/BC or the tangent of angle ACB. In Fig, 26 $AB = 82$ and $BC = 122$. Therefore, the **ratio of variation** equals $82/122 = 0.67$. This means that, on the average, for every change of 1 per cent. in the subject, there is a tendency for the relative to change in a like direction 67/100 of 1 per cent. The complement of this fraction, or 0.33, may be called the **ratio of regression.**

Sec. 123. **Elimination of Long-time Changes.**

Just as in the case of correlation so in computing the ratio of variation, long-time trends must be eliminated before we can study the relationship between the short-time oscillations of two historical variables but the process differs slightly. The first step is to decide certainly whether there is a trend of such a character that it may be brought out by a moving average. If not, the method, described in Sec. 121, of taking deviations from the arithmetic average only is preferable. When a definite trend exists and the wave-length of the short-time oscillations has been determined, the

next step is to compute the moving-average line for each series. The third operation is to divide each item by the moving average of that series for that year. This gives two series of index numbers which may be plotted in pairs exactly as were the indices in Sec. 122. A study of the following hypothetical table may help to make the method plain.

TABLE XXIII.

DATA FOR GALTON GRAPH TO SHOW THE RATIO OF VARIATION FOR SHORT-TIME CHANGES ONLY BETWEEN BANK RESERVES AND CHECK CIRCULATION.

Date.	Subject.			Relative.		
	B Bank Reserves in Dollars per Capita	b Moving Average.	$\frac{B}{b}$ Index.	C Check Circulation in Dollars per Capita.	c Moving Average.	$\frac{C}{c}$ Index.
1880	8.2			27		
1881	8.0			26		
1882	7.2	7.7	94	22	26	85
1883	7.4	7.6	97	25	27	93
1884	7.7	7.5	103	30	28	107
1885	7.7	7.5	103	32	29	110
1886	7.5	7.4	101	31	30	103
1887	7.2	7.3	99	27	31	87
1888	6.9	7.3	95	30	32	94
1889	7.2	7.3	99	35	33	106
1890	7.7	7.2	107	37	35	106
1891	7.5	7.1	106	36	37	97
1892	6.7	7.0	96	37	38	97
1893	6.4	6.9	93	40	39	103
1894	6.7	6.7	100	40	40	100
1895	7.2	6.6	109	42	40	105
1896	6.5	6.5	100	41	39	105
1897	6.2			37		
1898	5.9			35		

16

We observe that the bank reserves per capita show a constant tendency to diminish while, at the same time, the per capita check circulation is steadily increasing. But a change of $1 per capita does not represent the same **proportion** when the normal check circulation stands at $26, that it does when the normal circulation has increased to $39. Hence it is necessary in obtaining the correct **proportional** change to first obtain a moving average of each series. This gives a correct base at each point. If the original item is now divided by the moving average for that date, the resulting index shows correctly the **relative** increase or decrease due to the short-time influence. The pairs of indices may now be plotted in exactly the same manner as were those in Fig. 26.

Sec. 124. **The Correlation Table.**

When the items in both subject and relative are very numerous, it requires a large amount of effort to reduce each to an index number and plot it as was done in Fig. 26. To obviate this difficulty, which is especially important in the case of biological data, where the items may be extremely numerous, the **correlation table** has been invented. Its purpose is to group together the items of the subject in classes in a frequency table and then find some kind of an average of the items of the corresponding class of the relative and compare this average with the median of the class in the subject. It is, then, simply a method of substituting averages

for the individual pairs of items. The simplest form of the correlation table is given below, the subject being the lengths of a set of leaves, chosen at random, and the relative being the corresponding breadths of the same leaves. The first two columns under the heading "Subject" form a simple frequency table of the leaf lengths. To the right of each class of lengths under the title "Relative," we find a record of the breadth of each leaf belonging to that class of lengths. In the next to the last column of the table, we find the average breadth of all the leaves contained in that length class and it is this average which is to be compared with the median **length** of that class in order to obtain the ratio of variation. Reduction to index numbers is essential, for the same reasons heretofore described, before the data are ready for plotting as a Galton graph.

A few statements concerning the construction of the table may be helpful.

The class-intervals in the subject must all be equal and the same rule applies to the class-intervals within the relative. Classes covering all the data between the extremes of each group should be entered in the table even if no items occur in certain of the classes. In drafting a preliminary correlation table, it is customary, as the leaves are checked off in order of their length, to place a dot in the proper square to represent the breadth of each leaf. The dots are afterwards, simply counted and their sum entered in the corresponding square in the permanent table.

TABLE XXIV.

TABLE FOR CORRELATING THE LENGTHS AND BREADTHS OF LEAVES.

Subject.			Relative.											
Number of Leaves.	Length of Leaves in Mm.	Median Length of Class Divided by Average Length of All Leaves.	Breadth of Leaves in Mm.									Total Number.	Average Breadth of Leaves in Group.	Average Breadth of Class Divided by Average Breadth of All.
			13–15	16–18	19–21	23–25	26–28	29–31	32–34	35–37	38–40			
3	30–36	51.3	2	1								3	15.0	56.2
5	37–43	62.2	1	2	2							5	17.6	66.0
11	44–50	73.1		1	4	4	1	1				11	22.7	85.1
9	51–57	84.0			3	4	2					9	23.3	87.3
16	58–64	94.9		2	2	4	4	3	1			16	25.1	94.1
26	65–71	105.7		1	2	5	8	4	3	2	1	26	27.8	104.2
14	72–78	116.6					4	7	2	1		14	30.0	112.4
10	79–85	127.5				1	1	2	3	2	1	10	32.1	120.3
4	86–91	138.4							1	2	1	4	36.0	134.9
1	92–98	147.7									1	1	39.0	146.2
99		Totals	3	7	13	18	20	17	10	7	4	99		
Average	64.31											Average	26.67	

The column near the right-hand margin marked, "Total Number," is merely a check column and must correspond with the first column of the table.

The ideal average for use in all studies concerning the ratio of variation seems to be the mode, for it is the real center from which deviations take place. The fact, however, that the mode is, so commonly, ill defined, especially when the items are few in number, renders it less satisfactory as a basis than the arithmetic average, which has the merit of being absolutely definite. It is the latter, then, which is, in practice, most frequently employed and which forms the base in the model correlation table shown above.

Bowley suggests[1] the use of the median as being more convenient and just as accurate. It is one of the best averages when the items are numerous but not quite so satisfactory if they are badly scattered. If the class-intervals are large, it is necessary to interpolate, within the class, for the median item, according to the formula laid down in Sec. 71. If, however, the classes are very narrow it may be sufficiently accurate to consider the item at the midpoint of its class though this usually introduces a slight error.

When the arithmetic average is employed, as in Table XXIV, it is computed by assuming that the lengths or breadths of all the leaves within any class or sub-class fall at the midpoint of the same; thus, in the class 44–50,

[1] A. L. Bowley, *Elements of Statistics*, p. 323

the leaves are all considered as having a length of 47 mm. each. In the third column, we have the leaf lengths reduced to a series of index numbers by the process of dividing the midpoint of each class by the average length of all the leaves in the entire group. An index series is likewise obtained for the breadths by dividing the average breadth of leaves in each of the final classes in the relative by the average breadth of all the leaves.

A little study of the table will show that, in addition to its use in determining the ratio of variation, it gives considerable information concerning the correlation between the subject and the relative. As the leaves become longer, if there is correlation, they should become broader also. This will cause the modal breadth to shift constantly toward the right. The fact that the modes proceed regularly across the table at an angle approaching 45° indicates correlation. In the given table the line of modes slopes downward to the **right.** This indicates **direct** correlation. If they formed a regular slope falling toward the **left**, the correlation might be just as marked but would be **inverse**. The more closely the modes follow a straight line across the table and the closer the items are packed about the modes the higher is the degree of correlation indicated by the table. The line of totals at the foot of the table should, if there is correlation, show a well-defined mode. The summation of these totals acts as

another check on the correctness of the numbers of leaves entered in the table.

If one imagines the correlation table to be a plane surface and the numbers of items entered in the squares of the relative to represent altitudes at those respective points, he can picture to himself a sort of rugged hill with its crest near the center of the table. The surface of this hill is known as the correlation surface. If the correlation were perfect, the base of the hill would be oval in shape, the center of the hill would be in the center of the table, the surface would be smooth and it would slope off regularly in all directions so that a vertical section, cut through the crest in any direction, would always present the bell-shaped form of the normal frequency curve.

The correlation table being completed properly, the final step is to plot, in a Galton graph, the index of each class of the subject as an ordinate with the index of the corresponding class of the relative as the abscissa. The line of regression is drawn and the ratio of variation computed exactly as described in Sec. 122.

Sec. 125. **Conclusion.**

We have now finished, in addition to a brief review of the history of statistics, a discussion of the elementary methods most necessary in the study or manipulation of simple statistical data. We have covered the factors in the collection, analysis, and comparison of large numbers which are believed to be most essential to

the practical statistician, especially in the field of the social sciences. For a discussion of details, the development of the mathematical theory, or the practical applications of statistics the reader is referred to more advanced works.

REFERENCES.

GALTON, FRANCIS. *Correlations and their Measurement.* Proceedings of the Royal Society, 1888, XIV, 135.

BOWLEY, A. L. *Elements of Statistics*, pp. 322–326.

YULE, G. U. *Introduction to Statistics*, Chap. IX, and pp. 203–205.

PERSONS, WARREN M. *The Correlation of Economic Statistics.* Quar. Publications of the Amer. Statistical Assoc., Dec.,

APPENDIX A.

Calculating Devices.

The statistician will find that certain mechanical aids are essential if he is to do much actual statistical work. For operations in which it is not desired to attain greater accuracy than three digits, small slide-rules costing from $2.75 to $21.50 are very satisfactory. A Fuller spiral slide rule, price $30.00, is very convenient in multiplication and division and the results may be read correctly to five digits. The Thacher cylindrical slide rule costing $35.00 is slightly less convenient but has the additional merit of giving squares and square roots directly. Its accuracy is the same as that of the Fuller. If it is necessary to have exact results in multiplication or division, an arithmometer or reckoning machine is very desirable. These cost from $193 to $338. They may also be used for addition or subtraction, but, for the former purpose, are not so rapid as an ordinary adding machine. All of the above may be obtained of Keuffel, Esser and Co., of New York and Chicago. An adding machine is almost essential if a large amount of adding is to be done. The most complete ones are sold by the Burroughs Adding Machine Co., of Detroit, Mich., at from $325 up. The Wales adding machine manufactured by the Adder Machine Co., of Wilkes-Barre, Pa., is a slightly less expensive

machine which is satisfactory for most purposes. A still cheaper machine which adds perfectly but does not print is the Comptometer sold by the Felt and Tarrant Mfg. Co., of Chicago, Ill.

Where great numbers of items are to be tabulated at regular intervals, as in the case of expense accounts or pay rolls of municipalities or large companies, public utility reports, and the like, the Hollerith tabulating and sorting machines are great time savers. The items must first be punched on printed cards and the results are then automatically classified in any form or the various items of any group summated, as desired. These machines may be rented from the Hollerith Tabulating Machine Co., of Washington, D. C., at from $25 up, per month, each.

Many books of mathematical tables are published which are less expensive than machines and, for some purposes, are more satisfactory. The following are some of the best.

BARLOW'S *Tables of Squares, Cubes, Square-roots, Cube-roots, and Reciprocals of all Integer Numbers up to 10,000.* E. Spon, N. Y. Price. $2.50.

BOWSER, E. A. *Five-place Logarithmic Tables.* D. C. Heath & Co., Chicago. Price 50c.

BAUSCHINGER AND PETERS. *Logarithmic Tables.* Asher & Co., London. Gives eight-figure logarithms of numbers up to 200,000. Price 18s. 6d.

COTSWORTH, M. B. *The Direct Calculator.* Series 0. McCorquodale and Co., London. Gives products up to 1,000 × 1,000. Price, with index, 25s.

CRELLE, A. L. *Rechentafeln.* G. Reimer, Berlin. Gives products up to 1,000×1,000. Price 15m.

JONES, G. W. *Logarithmic Tables.* G. W. Jones, Ithaca, N. Y. Gives six-place logarithms of numbers up to 10,000, trigonometric functions, squares, cubes, roots, etc. Price $1.00.

LUDLOW, H. H. *Logarithmic and Other Mathematical Tables.* John Wiley & Sons, N. Y. Price $2.00.

PETERS, J. *Neue Rechentafeln für Multiplikation and Division.* G. Reimer, Berlin. Gives products up to 100×10,000. Can be obtained with English introduction. Price 15m.

VAN VELZER, C. A. *Four-Place Logarithmic and Trigonometric Tables.* Tracy, Gibbs & Co., Madison, Wis. Price 30c.

WELLS, W. *Six Place Logarithmic Tables.* D. C. Heath & Co., Chicago. Price 60c.

ZIMMERMANN, H. *Rechentafel.* Asher & Co., London. Gives products up to 100×1,000; also tables of squares, cubes, square roots, cube roots, etc. Price 5s.

APPENDIX B.

TABLE XXV.

Table of Logarithms of Numbers.

No.	0	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757

30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396

TABLE XXV.—*Continued.*

TABLE OF LOGARITHMS OF NUMBERS.

No.	0	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745

75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996

APPENDIX C.

TABLE XXVI.

TABLE OF SQUARES.

No.	0	1	2	3	4	5	6	7	8	9
10	10000	10201	10404	10609	10816	11025	11236	11449	11664	11881
11	12100	12321	12544	12769	12996	13225	13456	13689	13924	14161
12	14400	14641	14884	15129	15376	15625	15876	16129	16384	16641
13	16900	17161	17424	17689	17956	18225	18496	18769	19044	19321
14	19600	19881	20164	20449	20736	21025	21316	21609	21904	22201
15	22500	22801	23104	23409	23716	24025	24336	24649	24964	25281
16	25600	25921	26244	26569	26896	27225	27556	27889	28224	28561
17	28900	29241	29584	29929	30276	30625	30976	31329	31684	32041
18	32400	32761	33124	33489	33856	34225	34596	34969	35344	35721
19	36100	36481	36864	37249	37636	38025	38416	38809	39204	39601
20	40000	40401	40804	41209	41616	42025	42436	42849	43264	43681
21	44100	44521	44944	45369	45796	46225	46656	47089	47524	47961
22	48400	48841	49284	49729	50176	50625	51076	51529	51984	52441
23	52900	53361	53824	54289	54756	55225	55696	56169	56644	57121
24	57600	58081	58564	59049	59536	60025	60516	61009	61504	62001
25	62500	63001	63504	64009	64516	65025	65536	66049	66564	67081
26	67600	68121	68644	69169	69696	70225	70756	71289	71824	72361
27	72900	73441	73984	74529	75076	75625	76176	76729	77284	77841
28	78400	78961	79524	80089	80656	81225	81796	82369	82944	83521
29	84100	84681	85264	85849	86436	87025	87616	88209	88804	89401

30	90000	90601	91204	91809	92416	93025	93636	94249	94864	95481
31	96100	96721	97344	97969	98596	99225	99856	100489	101124	101761
32	102400	103041	103684	104329	104976	105625	106276	106929	107584	108241
33	108900	109561	110224	110889	111556	112225	112896	113569	114244	114921
34	115600	116281	116964	117649	118336	119025	119716	120409	121104	121801
35	122500	123201	123904	124609	125316	126025	126736	127449	128164	128881
36	129600	130321	131044	131769	132496	133225	133956	134689	135424	136161
37	136900	137641	138384	139129	139876	140625	141376	142129	142884	143641
38	144400	145161	145924	146689	147456	148225	148996	149769	150544	151321
39	152100	152881	153664	154449	155236	156025	156816	157609	158404	159201
40	160000	160801	161604	162409	163216	164025	164836	165649	166464	167281
41	168100	168921	169744	170569	171396	172225	173056	173889	174724	175561
42	176400	177241	178084	178929	179776	180625	181476	182329	183184	184041
43	184900	185761	186624	187489	188356	189225	190096	190969	191844	192721
44	193600	194481	195364	196249	197136	198025	198916	199809	200704	201601
45	202500	203401	204304	205209	206116	207025	207936	208849	209764	210681
46	211600	212521	213444	214369	215296	216225	217156	218089	219024	219961
47	220900	221841	222784	223729	224676	225625	226576	227529	228484	229441
48	230400	231361	232324	233289	234256	235225	236196	237169	238144	239121
49	240100	241081	242064	243049	244036	245025	246016	247009	248004	249001
50	250000	251001	252004	253009	254016	255025	256036	257049	258064	259081
51	260100	261121	262144	263169	264196	265225	266256	267289	268324	269361
52	270400	271441	272484	273529	274576	275625	276676	277729	278784	279841
53	280900	281961	283024	284089	285156	286225	287296	288369	289444	290521
54	291600	292681	293764	294849	295936	297025	298116	299209	300304	301401

TABLE XXVI.—*Continued.*

TABLE OF SQUARES.

No.	0	1	2	3	4	5	6	7	8	9
55	302500	303601	304704	305809	306916	308025	309136	310249	311364	312481
56	313600	314721	315844	316969	318096	319225	320356	321489	322624	323761
57	324900	326041	327184	328329	329476	330625	331776	332929	334084	335241
58	336400	337561	338724	339889	341056	342225	343396	344569	345744	346921
59	348100	349281	350464	351649	352836	354025	355216	356409	357604	358801
60	360000	361201	362404	363609	364816	366025	367236	368449	369664	370881
61	372100	373321	374544	375769	376996	378225	379456	380689	381924	383161
62	384400	385641	386884	388129	389376	390625	391876	393129	394384	395641
63	396900	398161	399424	400689	401956	403225	404496	405769	407044	408321
64	409600	410881	412164	413449	414736	416025	417316	418609	419904	421201
65	422500	423801	425104	426409	427716	429025	430336	431649	432964	434281
66	435600	436921	438244	439569	440896	442225	443556	444889	446224	447561
67	448900	450241	451584	452929	454276	455625	456976	458329	459684	461041
68	462400	463761	465124	466489	467856	469225	470596	471969	473344	474721
69	476100	477481	478864	480249	481636	483025	484416	485809	487204	488601
70	490000	491401	492804	494209	495616	497025	498436	499849	501264	502681
71	504100	505521	506944	508369	509796	511225	512656	514089	515524	516961
72	518400	519841	521284	522729	524176	525625	527076	528529	529984	531441
73	532900	534361	535824	537289	538756	540225	541696	543169	544644	546121
74	547600	549081	550564	552049	553536	555025	556516	558009	559504	561001

75	562500	564001	565504	567009	568516	570025	571536	573049	574564	576081
76	577600	579121	580644	582169	583696	585225	586756	588289	589824	591361
77	592900	594441	595984	597529	599076	600625	602176	603729	605284	606841
78	608400	609961	611524	613089	614656	616225	617796	619369	620944	622521
79	624100	625681	627264	628849	630436	632025	633616	635209	636804	638401
80	640000	641601	643204	644809	646416	648025	649636	651249	652864	654481
81	656100	657721	659344	660969	662596	664225	665856	667489	669124	670761
82	672400	674041	675684	677329	678976	680625	682276	683929	685584	687241
83	688900	690561	692224	693889	695556	697225	698896	700569	702244	703921
84	705600	707281	708964	710649	712336	714025	715716	717409	719104	720801
85	722500	724201	725904	727609	729316	731025	732736	734449	736164	737881
86	739600	741321	743044	744769	746496	748225	749956	751689	753424	755161
87	756900	758641	760384	762129	763876	765625	767376	769129	770884	772641
88	774400	776161	777924	779689	781456	783225	784996	786769	788544	790321
89	792100	793881	795664	797449	799236	801025	802816	804609	806404	808201
90	810000	811801	813604	815409	817216	819025	820836	822649	824464	826281
91	828100	829921	831744	833569	835396	837225	839056	840889	842724	844561
92	846400	848241	850084	851929	853776	855625	857476	859329	861184	863041
93	864900	866761	868624	870489	872356	874225	876096	877969	879844	881721
94	883600	885481	887364	889249	891136	893025	894916	896809	898704	900601
95	902500	904401	906304	908209	910116	912025	913936	915849	917764	919681
96	921600	923521	925444	927369	929296	931225	933156	935089	937024	938961
97	940900	942841	944784	946729	948676	950625	952576	954529	956484	958441
98	960400	962361	964324	966289	968256	970225	972196	974169	976144	978121
99	980100	982081	984064	986049	988036	990025	992016	994009	996004	998001

INDEX.

Numbers refer to sections, not pages.